Perforamances statistiques d'estimateurs non-linéaires

Michael Chichignoud

UNIVERSITÉ DE PROVENCE
U.F.R. M.I.M.
ÉCOLE DOCTORALE DE MATHÉMATIQUES ET INFORMATIQUE E.D. 184

THÈSE

présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE PROVENCE

Spécialité : Mathématiques

par

Michaël CHICHIGNOUD

sous la direction du Pr. Oleg LEPSKI

Titre :

Performances statistiques d'estimateurs non-linéaires

soutenue publiquement le 25 novembre 2010

JURY

M. Laurent CAVALIER	Université de Provence	Examinateur
Mme Béatrice LAURENT	INSA de Toulouse	Examinatrice
M. Oleg LEPSKI	Université de Provence	Directeur
M. Pascal MASSART	Université Paris-Sud	Rapporteur
Mme Dominique PICARD	Université Paris-Diderot	Examinatrice
M. Alexandre TSYBAKOV	Université Pierre et Marie Curie	Examinateur
M. Aad Van der VAART	Vrije Universiteit	Rapporteur

A mes grands parents,
Jacqueline et Robert,
une pensée pour eux.

REMERCIEMENTS

Mes premières pensées vont à mon directeur de thèse, Oleg Lepski, qui m'a fait l'immense honneur de m'accompagner pendant ces trois longues années. Il est difficile en quelques lignes d'exprimer tous mes remerciements et mon estime pour lui. J'ai souvent mis à contribution ses nombreuses qualités humaines, mais son soutien et ses conseils avisés m'ont permis de réaliser ce manuscrit. Par ses connaissances scientifiques et son extraordinaire niveau en mathématiques, qui m'impressionne toujours, il a su me lancer sur un sujet très ouvert et je le remercie très sincèrement. Dans les années futures, j'espère pouvoir répondre aux nombreuses questions ouvertes, qu'il m'a posé (et auxquelles je n'ai pas su répondre), sur des problèmes qui occuperont certainement une place importante dans les statistiques de demain.

J'exprime tous mes remerciements à Pascal Massart et Aad Van der Vaart pour avoir gentiment accepté de rapporter cette thèse. Je suis très touché qu'ils aient pris le temps de relire mes modestes travaux.

Je tiens également à remercier Laurent Cavalier, Béatrice Laurent, Dominique Picard et Alexandre Tsybakov, de l'immense honneur qu'ils me font de participer à ce jury de thèse d'une très grande qualité scientifique.

Un grand merci à tous les membres de l'équipe Probabilités/Statistiques, notamment à Florent Autin, Yuri Golubev, Grégory Maillard, Christophe Pouet et Thomas Willer pour leurs conseils avisés, leur sympathie et leur engouement pour les mathématiques.

Je voudrais remercier aussi les personnes qui ont contribué à la réalisation de cet ouvrage, notamment Christophe Pouet et Thomas Willer. Ma sincère gratitude et mon amitié vont à Joseph Salmon, futur docteur lui aussi. Un petit clin d'oeil à Adrien Saumard, qui a su, en peu de temps, me poser des questions très ouvertes sur mon sujet et me donner quelques conseils de rédaction.

Je tiens à remercier les enseignants qui m'ont le plus marqué au cours de mes études, et qui m'ont fait aimer les mathématiques : M. Fournier (Professeur de mathématiques en Première) pour son enthousiasme et son dé à *28 faces*, Dominique Barbolosi (Professeur en Licence) pour la qualité de ses cours sans support papier, et enfin Laurent Cavalier (Professeur en Master) pour ses cours dynamiques effectués avec beaucoup de rigueur.

Ces trois années passées au CMI m'ont permis de rencontrer des gens extraordinaires. Je remercie le personnel administratif pour leur disponibilité et leur sympathie, les charges

administratives étaient bien peu lourdes grâce à leur présence. Je suis heureux d'avoir fait partie de la Team des doctorants du Cmi et je remercie chacun d'entre eux pour les nombreuses discussions que nous avons eu sur divers domaines des mathématiques. Je remercie sincèrement Hamish Short (directeur de l'école doctorale) et Etienne Pardoux (ex-directeur de l'école doctorale) pour tout ce qu'ils ont fait pour moi, en particulier pour l'obtention d'une bourse doctorale. Un grand merci aux membres du bureau 114 que j'ai pu côtoyés au quotidien, notamment Clément Marteau, Sébastien Loustau et Shanti Gibert. Leur présence fût chaleureuse et indispensable durant toutes ces années.

J'exprime mes plus sincères remerciements à ma famille pour son accompagnement dans ce long parcours. Ma mère, mon frère et mes grands parents ont été d'un soutien sans faille. En particulier, un immense merci à mon père qui à force de patiente et d'encouragements, m'a communiqué son penchant pour les mathématiques et la logique. Du fond du coeur Merci !

Et pour finir, tous mes sentiments vont à ma future femme Mylène, qui a accepté de me dire "oui" pour me rendre le plus heureux des hommes. Sans elle et son chaleureux soutien, je ne serais pas là aujourd'hui. Merci mon amour !

SOMMAIRE

Le thème de cette thèse est l'estimation non-paramétrique, en particulier l'étude des performances théoriques de l'estimation de fonctions de régression, une partie importante d'un des domaines mathématiques connue sous le nom de *statistique mathématique*.

Une partie de cette thèse est rédigée en anglais, car ce sont des preprints ou des articles soumis. Les chapitres 1 et 2 sont quant à eux rédigés en français. C'est pourquoi nous avons fait précéder chaque chapitre d'un court résumé en français.

Ce manuscrit est organisé comme suit :

- Le chapitre 1 est une introduction aux modèles de régression, aux critères de performance en vigueur et aux méthodes adaptatives avec sélection de fenêtres.
- Dans le chapitre 2, le lecteur peut trouver les principaux résultats de cette thèse pour deux nouveaux types d'estimateurs. Notons qu'un lecteur non-spécialiste du domaine aura recours au chapitre 1 pour sa compréhension.
  Nous proposons aussi quelques perspectives à ce travail. Notamment, une liste d'une quinzaine de problèmes ouverts est donnée à la fin de ce chapitre.
- Les chapitres 3 et 4 traitent de l'estimation bayésienne.
- Le chapitre 5 introduit l'estimateur de Huber et quelques résultats sur ses performances.

Les résultats théoriques sont accompagnés d'expériences numériques. En particulier, on pourra comparer les estimateurs *bayésien* et de *Huber* avec les estimateurs linéaires.

Les chapitres 3 à 5 peuvent être lus indépendamment des autres (ce qui est à l'origine de quelques répétitions). Il reste néanmoins quelques liens (limités aux outils probabilistes et à des résultats techniques communs). Nous avons fait tout notre possible pour uniformiser nos notations, que nous définissons avant le chapitre 3 et qui sont redéfinies ensuite pour chaque chapitre.

La version électronique de cette thèse (à télécharger sur la page web de l'auteur : http ://www.latp.univ-mrs.fr/∼chichign/doku.php ?id=acceuil) comprend des liens *hyperref* qui permettent de se retrouver au chapitre, à la section, à la formule ou à la citation en un simple clic sur la référence. Nous incitons le lecteur à l'utiliser pour une recherche

bibliographique ou un résultat technique, bien que la version papier reste la plus agréable à lire.

Table des matières

Chapitre 1

Régression Non-Paramétrique

Nous présentons, dans ce chapitre introductif, les modèles de régression non-paramétriques que nous étudions. Nous introduirons l'approche localement paramétrique, le risque minimax et la notion d'adaptation. Les estimateurs, utilisés dans cette thèse, sont présentés dans la section 1.3. Une partie importante de l'introduction est consacrée à une présentation détaillée de la méthode dite de *Lepski* (voir Section 1.5.2).

1.1 Objet de la Thèse

Dans cette thèse, nous étudions un domaine de la statistique mathématique : *l'estimation non-paramétrique*. Ceci consiste à estimer des fonctions (objets de dimensions infinies) à partir d'observations "bruitées". Ce genre d'approche s'est considérablement développé ces dernières années dans le monde scientifique. L'imagerie (médicale ou astronomique), l'étude du génome (puces à ADN en grande dimension) ou encore les problèmes inverses (physique des matériaux, tomographie en imagerie médicale, etc.) ont recours à l'estimation non-paramétrique. Les statistiques ont un attrait particulier du fait qu'elles utilisent des théories mathématiques pour modéliser des problèmes réels. En particulier, l'introduction, de la notion d'*aléa* (en anglais : *random*) dans les observations, permet au statisticien de construire des méthodes d'estimation "fiables" en théorie.

La modélisation de problème se fait de la manière suivante. On dispose d'un nombre n de données, notées $Y = (Y_1, Y_2, ..., Y_n) \in \mathbb{R}^n$, issues d'une expérience réelle. On modélise les observations en utilisant deux théories mathématiques, l'analyse fonctionnelle et les probabilités. On peut décomposer les observations Y de la façon suivante :

$$Y_i = f(X_i) + \text{"bruit"}, \quad i = 1, ..., n, \quad X_i \in \mathbb{R}^d,$$

que l'on appele *modèle additif*, où f est une fonction "régulière" à d variables de $\mathbb{R}^d \to \mathbb{R}$. Nous supposerons pour toute cette thèse que la fonction f est dans un certain espace fonctionnel noté $\mathcal{F}_\beta$ de dimension infinie où β est un paramètre de régularité.

On suppose également que nos observations sont dégradées par un aléa inconnu que nous modélisons par une variable aléatoire comme nous le voyons souvent en statistique (voir Section 1.2).

A partir de ce modèle mathématique, on veut résoudre le problème suivant : Estimer ou reconstruire la fonction f à partir des observations $(Y_1, Y_2, ..., Y_n)$. Pour cela nous développons des outils mathématiques adéquats. Dans les modèles additifs, comme le modèle gaussien, des méthodes d'estimation ont été largement développées depuis un demi-siècle et sont devenus un outil courant de l'estimation non-paramétrique : estimateurs à noyau (voir Rosenblatt [1956], Parzen [1962], Nadaraya [1964], Watson [1964] et Borovkov [1987]) ou plus récemment la décomposition en bases d'ondelettes (Donoho, Johnstone, Kerkyacharian, et Picard [1995] et Härdle, Kerkyacharian, Picard, et Tsybakov [1998]).

Il est bien connu que les estimateurs linéaires (par rapport aux observations, voir Définition 4) ne sont pas *robustes* (non-sensibles aux valeurs extrêmes, en anglais *outliers*). Par exemple, si le bruit est une variable aléatoire de Cauchy (sans moment d'ordre 1), ces estimateurs sont inefficaces (Voir Exemple 1.1).

Si les observations admettent un moment d'ordre deux ($\mathbb{E}Y^2 < \infty$), on sait, d'après le théorème central limite, que la moyenne empirique converge (en probabilité) vers son espérance à la vitesse $1/\sqrt{n}$. Ainsi, les estimateurs linéaires ne feront jamais mieux que cette vitesse. Ce qui implique que les estimateurs linéaires ne sont pas optimaux pour certains modèles (voir Exemple 1.1 et Section 1.2).

L'utilisation du phénomène de Stein [1981] repose sur le fait que le bruit est gaussien. En effet, Stein [1981] met au évidence la contraction de Stein dans le modèle de suites gaussiennes qui n'est valable seulement si le bruit est gaussien ou sous-gaussien. Bien que l'estimateur de James/Stein soit "meilleur" que la moyenne empirique, ces propriètés théoriques sont valables seulement dans le cas Gaussien.

L'adaptation, en estimation non-paramétrique, est aussi très "gourmande" en bruit gaussien, notamment pour l'obtention d'inégalités exponentielles de concentration. Par exemple pour l'estimateur à noyau, ceci est nécessaire (voir Section 1.5).

Exemple 1. Dans le cas paramétrique, on peut trouver des modèles dans lesquels les estimateurs linéaires ne sont pas optimaux. Par exemple, si les observations suivent une loi uniforme continue sur l'intervalle $[0, \theta]$, $Y_i \sim \mathcal{U}_{[0,\theta]}$, $i = 1, ..., n$, alors la moyenne empirique notée $\bar{Y}$ permet d'estimer $\theta \geq 0$ avec la vitesse $1/\sqrt{n}$,

$$2\bar{Y} \xrightarrow[n\to\infty]{n^{-1/2}} \theta, \quad \text{en probabilité.}$$

On peut construire un estimateur plus rapide dans ce modèle,

$$\max_i Y_i \xrightarrow[n\to\infty]{n^{-1}} \theta, \quad \text{en probabilité,}$$

un estimateur *non-linéaire* qui atteint la vitesse $1/n$. Nous étudions certains modèles avec cette particularité dans le cas non-paramétrique et nous développons un estimateur *bayésien*

non-linéaire. Celui-ci nous permet d'atteindre les vitesses de convergence optimales (Voir Définition de l'optimalité, Section 1.4).

Exemple 2. On regarde maintenant le modèle de Cauchy, pour lequel, beaucoup d'estimateurs ont échoué, notamment les estimateurs linéaires. On prend les variables $Y_i = \theta + C_i$, $i = 1, ..., n$ avec C_i une variable de Cauchy (Définition 1.2.4). Il est facile de voir que la moyenne empirique $\bar{Y}$ n'est pas convergente. Dans ce cas, on utilise la médiane empirique $\mathrm{Med}(Y)$ et sa normalité asymptotique (voir par exemple Brown, Cai, et Zhou [2008]). Ainsi, la médiane empirique converge vers θ à la vitesse $1/\sqrt{n}$,

$$\mathrm{Med}(Y) \xrightarrow[n\to\infty]{n^{-1/2}} \theta, \quad \text{en probabilité.}$$

Pour le cas non-paramétrique, nous proposons un estimateur fondé sur cette idée que nous appelons *estimateur de Huber*.

Nous venons de présenter deux exemples où les estimateurs linéaires ne sont pas efficaces (voir inutilisables). Ces remarques motivent le travail de cette thèse avec l'introduction de nouveaux estimateurs non-linéaires.

Dans cette thèse, nous développons de nouveaux estimateurs localement paramétriques capables (sous certaines conditions) de s'adapter aux différents bruits (gaussien, Cauchy, etc.) et à la forme du modèle (additif ou multiplicatif). Ainsi, nous présentons deux types d'estimateurs : *estimateur bayésien* et *estimateur de Huber*. Nous montrons que pour différents modèles de régression, ces estimateurs sont optimaux au sens *minimax* (Définition 8) sur les *espaces de Hölder isotropes* (Définition 1).

La notion d'*adaptation*, introduite depuis une vingtaine d'années, est un point incontournable en estimation non-paramétrique (Stone [1982] et Efromovich et Pinsker [1984]). Nous proposons plusieurs procédures adaptatives reposant sur la méthode dite de Lepski pour le choix de la fenêtre. Ces procédures permettent aux estimateurs considérés d'atteindre des vitesses de convergences adaptatives optimales en un certain sens (voir Section 4.3). L'utilisation de la méthode dite de Lepski est fréquente dans cette thèse. De ce fait, nous présentons dans ce chapitre, l'idée de la méthode ainsi que les conditions suffisantes à la mise en oeuvre de la procédure.

L'intérêt de l'estimation dépasse largement le cadre pratique et algorithmique. Il y a un vrai intérêt à développer des procédures qui sont optimales en théorie et de pouvoir les utiliser dans différents domaines d'applications. Ainsi, les performances théoriques de nos estimateurs adaptatifs reposent sur le contrôle des *grandes déviations*. Ce contrôle peut être obtenu par des inégalités de type *inégalités de concentration* pour les processus empiriques qui ont été très développées, entre autres par Talagrand [1995,1996a,1996b], Ledoux [1997], Birgé et Massart [1998],
Massart [2000, 2007], Bousquet [2002], Boucheron, Bousquet, et Lugosi [2004], Golubev et Spokoiny [2009] et Goldenshluger et Lepski [2009b], du fait de leur nécessité.

1.2 Modèles de Régression et Espaces Fonctionnels

Dans cette section, nous présentons les espaces fonctionnels utilisés (espaces de Hölder), ainsi que les différents modèles de régression étudiés dans cette thèse. Dans un premier temps, nous parlerons du modèle de *régression générale* , et différents exemples de celui-ci, comme la régression additive avec densité inconnue, la régression gaussienne, la régression α et la régression avec bruit multiplicatif uniforme. Un rapide survol est donné sur les modèles des statistiques non-paramétriques en fin de section.

1.2.1 Espaces de Hölder Isotropes

Tout au long de ce manuscrit, nous travaillerons exclusivement avec les espaces de Hölder isotropes, i.e. la fonction de régression inconnue est supposée à plusieurs variables avec la même régularité hölderienne dans chaque direction (par rapport à chaque variable).

Pour tout $(p_1, ..., p_d) \in \mathbb{N}^d$ nous notons $\vec{p} = (p_1, ..., p_d)$ et $|\vec{p}| = p_1 + ... + p_d$.

Définition 1. *Soient $\beta > 0$, $L > 0$, $M > 0$, $d \in \mathbb{N}^*$ et $\lfloor \beta \rfloor$ le plus grand entier strictement inférieur à β. La* classe de Hölder isotrope $\mathbb{H}_d(\beta, L, M)$ *est l'ensemble des fonctions $f : [0,1]^d \to \mathbb{R}$ ayant sur $[0,1]^d$ toutes ses dérivées d'ordre $\lfloor \beta \rfloor$ et telles que $\forall x, y \in [0,1]^d$*

$$\sum_{m=0}^{\lfloor \beta \rfloor} \sum_{|\vec{p}|=m} \sup_{x \in [0,1]^d} \left| \frac{\partial^{|\vec{p}|} f(x)}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} \right| \leq M,$$

$$\left| \frac{\partial^{|\vec{p}|} f(x)}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} - \frac{\partial^{|\vec{p}|} f(y)}{\partial y_1^{p_1} \cdots \partial y_d^{p_d}} \right| \leq L \|x - y\|_1^{\beta - \lfloor \beta \rfloor}, \quad \forall \, |\vec{p}| = \lfloor \beta \rfloor.$$

La théorie Minimax repose essentiellement sur le fait que la fonction à estimer est dans un espace fonctionnnel, ici $\mathbb{H}_d(\beta, L, M)$. Les résultats minimax présentés dans la suite peuvent être étendus aux espaces anisotropes. Mais nous verrons dans la section 2.4 que les méthodes adaptatives ne sont pas conçues pour les espaces anisotropes dans le cas où l'estimateur est non-linéaire.

Pour étudier les modèles de régression inhomogène de Poisson et de régression multiplicative uniforme, une contrainte supplémentaire doit être rajoutée sur le support de $f \in \mathbb{H}_d(\beta, L, M, A)$ où

$$\mathbb{H}_d(\beta, L, M, A) = \left\{ f \in \mathbb{H}_d(\beta, L, M) \ : \ \inf_{x \in [0,1]^d} f(x) \geq A \right\}, \quad A > 0.$$

Remarque 1. *Cette hypothèse est suffisante pour l'utilisation de l'estimateur bayésien pour les régressions inhomogène de Poisson et multiplicative uniforme, celle-ci est purement théorique (nécessaire) et n'est pas justifiée en pratique. Si on prend la régression multiplicative uniforme (définie dans la section 1.2.7), et si f admet des points de nullité, alors*

$Y_i = 0$. *Avec cette observation, il est difficile de dire si cela est dû à la fonction ou au bruit (qui a une probabilité non-négligeable d'être proche de 0). Pour la régression inhomogène de Poisson, le paramètre de la loi de Poisson est toujours positif. Ainsi en pratique, on prendra A proche de* 0 *pour minimiser cette restriction. Sans cette restriction, le problème est ouvert.*

La classe des fonctions de Hölder est incluse dans les espaces de Besov qui sont souvent utilisés dans l'approche maxiset ou avec les estimateurs par ondelettes (Autin [2004]). Nous donnons la définition de ces espaces ci-dessous.

Les espaces de Besov Multidimensionnels. Nous rappelons la définition des *espaces de Besov* pour les fonctions unidimensionnelles et quelques inclusions avec les espaces de Hölder et Sobolev. On définit les espaces de Besov, dans le cas de fonctions de $[0,1]^d$ dans $\mathbb{R}$. Il faut pour cela caratériser le *module de continuité* d'une fonction f de $L^p\big([0,1]^d\big)$. Pour tout x dans $[0,1]^d$, notons $\Delta_h f(x) = f(x-h) - f(x)$, et pour tout entier u, on note l'itérée $\Delta_h^u f = \Delta_h \circ \cdots \circ \Delta_h f$. On définit alors le u^{e} *module de continuité* pour la norme p (avec $p \in [1,\infty]$) et pour tout $t > 0$ de la manière suivante

$$\omega^p(f,t) = \sup_{\|h\|_2 \le t} \left(\int_{J_{u,h}} |\Delta_h^u f(x)|^p dx \right)^{1/p},$$

où $J_{u,h} = \{x \in [0,1]^d, x + uh \in [0,1]^d\}$ et $\|.\|_2$ est la norme ℓ_2 sur $\mathbb{R}^d$.

Définition 2. *Soient $p \in [1,\infty]$, $q \in [1,\infty]$, $s \in]0,\infty[$ et $u = \lceil s \rceil$ (où $\lceil s \rceil$ est le plus petit entier strictement plus grand que s). On dit qu'une fonction f appartenant à $L^p\big([0,1]^d\big)$ est dans l'espace de Besov $\mathcal{B}^s_{p,q}\big([0,1]^d\big)$, quand $\gamma_{spq}(f) < \infty$ où*

$$\gamma_{spq}(f) = \begin{cases} \displaystyle\int_0^\infty \big(t^{-s}\omega^p(f,t)\big)^q \frac{dt}{t}, & \text{si } 1 \le q < \infty\,, \\ \displaystyle\sup_t |t^{-s}\omega^p(f,t)|, & \text{si } q = \infty\,. \end{cases}$$

On note $\mathcal{B}^s_{p,q}(L)$ une boule de Besov de rayon L, munie de la norme :

$$\|f\|_{\mathcal{J}^s_{p,q}} = \|f\|_p + \gamma_{spq}(f).$$

Les espaces de Besov constituent une très grande famille de fonctions. En particulier, rappelons que l'espace de Sobolev S^s correspond précisément à l'espace $\mathcal{B}^s_{2,2}$ et l'espace de Hölder $\mathbb{H}_d(s,L)$ (avec $0 < s \notin \mathbb{N}$) à l'espace $\mathcal{B}^s_{\infty,\infty}(L)$ où L est le rayon de la boule de Besov. Ces espaces sont très utilisés dans l'approche par ondelettes. En effet, il est possible de définir les espaces de Besov à partir des coefficients dans la base d'ondelettes (voir Härdle, Kerkyacharian, Picard, et Tsybakov [1998]). Pour plus de détails sur ces espaces, on se référera aux travaux de Bergh et Löfström [1976], Peetre [1976], Meyer [1992] ou DeVore et Lorentz [1993].

1.2.2 Régression Générale

Dans notre modèle, on observe les couples de variables aléatoires indépendantes (X_1, Y_1) , ..., (X_n, Y_n) notées

$$\mathcal{Z}_n = \big(X_i, Y_i\big)_{i=1,\dots,n}, \tag{1.2.1}$$

où X_i est un vecteur dit de variables explicatives (appelé *design*) qui détermine la distribution de l'observation Y_i. Le vecteur $X_i \in [0,1]^d$ de dimension d peut être vu comme une variable temporelle ou spatiale et $Y_i \in \mathbb{R}$ l'observation au point X_i. Notre modèle suppose que les valeurs X_i peuvent être aléatoires ou fixées et que la distribution de chaque Y_i est déterminée par un paramètre f_i qui peut dépendre de la position X_i, $f_i = f(X_i)$. Dans beaucoup de cas, la paramétrisation naturelle est choisie de la façon suivante $f_i = \mathbb{E}\big(Y_i|X_i\big)$ ($\mathbb{E}$ est l'espérance mathématique). On note $g(., f_i)$ la densité sur $\mathbb{R}$ de l'observation Y_i par rapport à la mesure de Lebesgue. Pour la *régression additive*, nous considèrerons le design aléatoire de loi uniforme sur $[0,1]^d$. Le problème d'estimation est de reconstruire la fonction f en tout point y. Ce modèle que nous appelons *Régression générale*, sera traité dans le Chapitre 3.

Un estimateur dit *bayésien* est construit pour ce modèle très général. Nous verrons que sous certaines conditions (voir hypothèses 3), cet estimateur atteint les vitesses de convergence adaptatives optimales. Dans cette thèse, cette approche introduite par Has'minskii et Ibragimov [1981], pour l'estimation paramétrique, est généralisée à l'adaptation. Certains modèles, où la densité g des observations est discontinue, sont étudiés.

1.2.3 Régression Additive

Dans ce modèle, on observe les couples de variables aléatoires indépendantes $(X_i, Y_i)_i$ qui vérifient l'équation :

$$Y_i = f(X_i) + \xi_i, \quad i = 1, ..., n, \tag{1.2.2}$$

où le bruit est une variable aléatoire de densité $g_\xi(.)$. X_i est le design aléatoire de loi uniforme sur $[0,1]^d$ et X_i est indépendant de ξ_i. La fonction de régression f est supposée appartenir à une boule de Hölder $\mathbb{H}_d(\beta, L, M)$. Nous supposerons vérifiées les hypothèses suivantes :

Hypothèses 1.

1. *g_ξ est paire,*
2. *$g_\xi(0) \geq A > 0$,*
3. *g_ξ est continue en 0.*

Remarque 2. *La symétrie du bruit (hypothèse 1.1) est une supposition assez faible. En général, on suppose que les effets du bruit sont les mêmes à gauche et à droite de la médiane qui vaut 0. Notons qu'un grand nombre de densités, utilisées dans cette thèse, vérifient cette*

hypothèse. Les deux autres hypothèses sont nécessaires pour contrôler les grandes déviations de l'estimateur de Huber (voir Chapitre 5). Elles permettent aussi de vérifier l'assertion suivante. Avec ces hypothèses, on peut voir que ξ admet une médiane théorique unique égale à 0, i.e. 0 est unique solution de l'équation $\mathbb{P}(\xi < x) = 1/2$. Noter qu'il existe de nombreux modèles classiques vérifiant ces hypothèses (Régressions gaussienne, de Cauchy et α). On peut remarquer que la régression additive est un cas particulier de la régression générale. En effet, il suffit de prendre $g\big(., f(X_i)\big) = g_\xi(. - f(X_i))$.

L'étude de ce modèle nécessite la construction d'un estimateur fondé sur l'idée de la médiane (estimateur de Huber, voir Chapitre 5). En particulier, nous développons une procédure adaptative dans le cas où la régularité de la fonction cible est inconnue.

Hall et Jones [1990] ont utilisé une méthode de *Validation croisée* (voir Tsybakov [2008]) sur une famille d'estimateurs robustes pour obtenir des résultats adaptatifs avec le risque L_2. Peu après, Härdle et Tsybakov [1992] ont étandu ce résultat, avec une méthode de *Plug-in* (introduit par Woodroofe [1970]), pour des *fonctions de contrast* plus générales, avec un choix aléatoire local de la fenêtre, mais seulement des résuultats de normalité asymptotique sont donnés.

L'adaptation dans la régression additive a fait l'objet d'un autre travail. En effet, Brown, Cai, et Zhou [2008] utilisent la normalité asymptotique de la médiane pour approximer ce modèle par le modèle gaussien avec une méthode de médiane par blocs. Une étape intermédiaire est de projeter les nouvelles observations dans une base d'ondelettes. Ensuite, la méthode de Stein par blocs (voir Cai [1999]) est utilisée pour l'adaptation en estimation globale. En revanche, cette approche nécessite des hypothèses plus fortes que les hypothèses 1 qui sont suffisantes pour l'utilisation de l'estimateur de Huber.

Plus récemment, Reiss, Rozenholc, et Cuenod [2009] utilise la méthode de Lepski pour développer un estimateur de Huber adaptatif dans le modèle additif. Les résultats obtenus sont pour l'estimation ponctuelle mais pour les fonctions localement constante (i.e. $\beta \leq 1$).

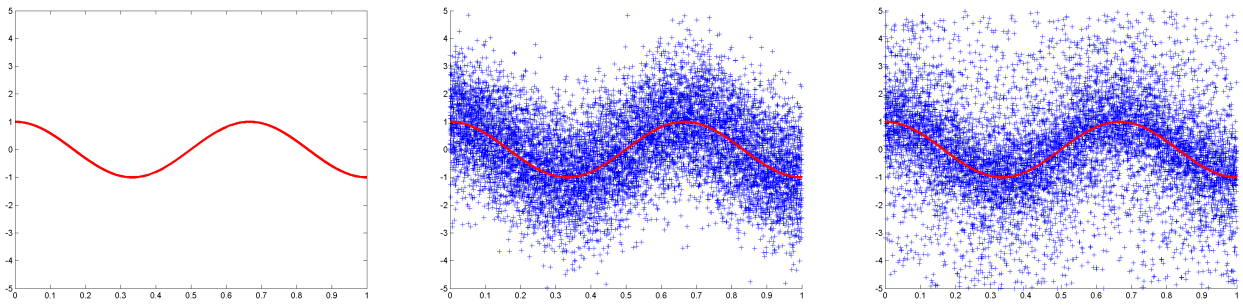
Dans la suite, nous présentons plusieurs cas particuliers où les estimateurs bayésien et de Huber peuvent être utilisés. En effet nous vérifions dans le cas de la régression gaussienne, si nos estimateurs sont applicables. Pour le bruit de Cauchy, l'estimateur de Huber est tout désigné pour estimer la fonction de régression (voir Chapitre 5). Nous présentons des modèles peu connus dans la littérature (voir Régressions α et multiplicative uniforme). Nous utilisons l'estimateur *bayésien* dans le cadre du modèle inhomogène de Poisson (très utilisé en imagerie, voir notamment Polzehl et Spokoiny [2006] et Katkovnik et Spokoiny [2008]). Ici la vitesse de convergence est la même que pour le modèle gaussien (qui peut être obtenu avec un estimateur linéaire). Les régressions, dites α et *multiplicative uniforme*, sont de très bons exemples dans lesquels la vitesse de convergence devient meilleure que la vitesse des estimateurs linéaires. Pour plus de détails, voir Section 2.1.4.

1.2.4 Régression Gaussienne et de Cauchy

Les régressions *gaussienne* ou de *Cauchy* sont des cas particuliers de la *régression additive* où la densité $g_\xi(.)$ prend respectivement les formes suivantes

$$\text{Densité gaussienne :} \quad g_\xi(x) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{x^2}{2\sigma^2}\right\}, \quad \sigma > 0,$$

$$\text{Densité de Cauchy :} \quad g_\xi(x) = \frac{1}{\pi}\frac{a}{a^2+x^2}, \quad a > 0.$$



(a) Fonction de régression. (b) Régression Gaussienne (c) Régression de Cauchy.

FIGURE 1.1 – Observations dans les régressions gaussiennes et de cauchy.

Néanmoins, elles suscitent un intérêt particulier. En effet le bruit gaussien est utilisé de façon systématique dans le domaine des statistiques et le bruit de Cauchy est une variable aléatoire sans moment d'ordre 1. Pour le bruit gaussien, on peut trouver une multitude d'articles qui l'étudient. Citons deux livres de Nemirovski [2000] et Tsybakov [2008] pour une introduction à ce modèle et aux méthodes classiques d'estimation non-paramétrique. Nous introduisons ce modèle standard pour vérifier que les estimateurs développés dans cette thèse fonctionnent correctement dans ce modèle classique (voir Chapitres 3 et 5).

L'estimateur de *Huber* est particulièrement bien adapté aux bruits dont les densités sont à queues lourdes (par exemple bruit de Cauchy, voir Chapitre 5). On peut constater une différence notable, entre les bruits gaussien et de Cauchy, dans la figure 1.1. En effet pour le bruit de Cauchy, on constate un nombre plus important de *valeurs extrêmes*. Ceci explique l'inefficacité des estimateurs linéaires pour ce bruit.

1.2.5 Régression Inhomogène de Poisson

Considérons la régression générale, cette fois avec la particularité que les observations sont discrètes $Y_i \in \mathbb{N}$. Nous supposons que Y_i suit une loi de Poisson de paramètre $f(X_i)$

$(Y_i \sim \mathcal{P}\big(f(X_i)\big))$. On écrit la densité

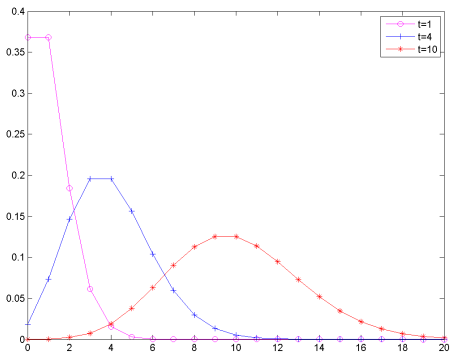
$$g(k, f(X_i)) = \mathbb{P}_f(Y_i = k) = \frac{\big[f(X_i)\big]^k}{k!} \exp\big\{ - f(X_i)\big\}, \quad k \in \mathbb{N}, \quad f \in \mathbb{H}_d(\beta, L, M, A).$$

Les points du design $(X_i)_{i\in 1,\dots,n}$ sont déterministes et sans perte de généralité nous supposons qu'ils sont répartis de manière uniforme sur la grille suivante :

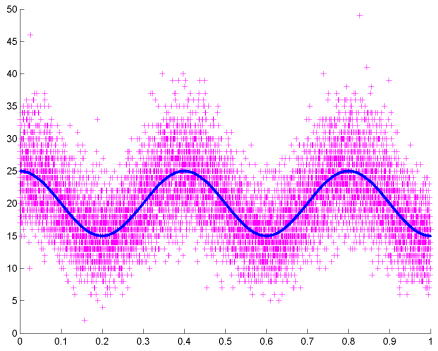
$$X_i \in \big\{1/n^{1/d}, 2/n^{1/d}, \dots, 1\big\}^d, \quad i = 1, \dots, n. \tag{1.2.3}$$

Remarquons la restriction de f à l'espace $\mathbb{H}_d(\beta, L, M, A)$, cette restriction ($f \geqq A$) est retrouvée dans la régression multiplicative uniforme. Ce modèle est très utilisé en imagerie, en particulier pour modéliser la photométrie des appareils photo numérique. On le trouve aussi en tomographie, il modélise la résonance magnétique des positons utilisée pour obtenir des IRM d'une ou plusieurs parties du corps humain.

Les travaux de Anscombe [1948] permettent de passer approximativement de la régression de Poisson à la gaussienne par une transformation des observations par la fonction $x \to 2\sqrt{x + 3/8}$. Dans la figure 1.2, nous donnons des exemples de densités de Poisson et un échantillon d'observations simulé à partir d'une fonction de régression.



(a) Densités d'une loi de Poisson de paramètre t.



(b) Observations de Poisson.

FIGURE 1.2 – Densité de la loi de Poisson de paramètre $t > 0$ et observations.

Image Originale.

Bruit de Poisson.

Bruit Gaussien.

FIGURE 1.3 – Illustration d'images issues d'observations de Poisson (zoomer sur la version numérique).

1.2.6 Régression α

Le modèle de régression α est le suivant

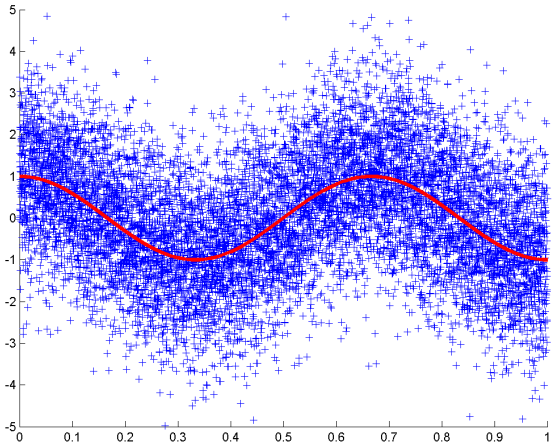
$$Y_i = f(X_i) + \epsilon_i, \quad i = 1, ..., n, \tag{1.2.4}$$

où ϵ_i a pour densité $g_\alpha(x) = C(\alpha) \exp\{-|x|^\alpha\}$ avec $0 < \alpha < 1/2$. $C(\alpha)$ est une constante choisie pour que g_α soit bien une densité. Ici le design X_i est choisi comme dans (1.2.3) et $f \in \mathbb{H}_d(\beta, L, M)$. La particularité de ce modèle est que pour $\alpha < 1/2$ la vitesse de convergence change en fonction de α (voir Chapitre 3).

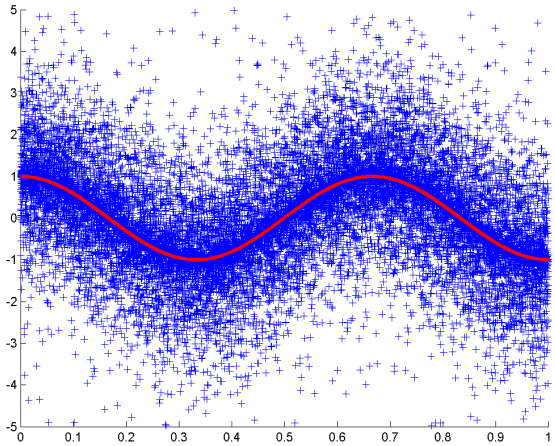
La figure 1.4 présente les observations de ce modèle pour $\alpha = 1/4$. On constate qu'il y a un très grand nombre d'observations éloignées de la fonction de régression f. Mais un petit nombre est très proche de f (à comparer avec la figure 1.1), cela nous donne une information très précieuse qui peut se vérifier en théorie à travers de meilleures vitesses de convergence. En effet, nous verrons que l'estimateur bayésien est plus rapide (au sens des vitesses de convergence) que les estimateurs linéaires (voir Sections 2.1 et 3.4.3).

Nous donnons la figure 1.5 pour illustrer ce comportement. En effet, la queue de distribution ne joue pas un rôle majeur entre les densités. Le comportement de la densité, à mettre en avant, est la concentration du bruit au voisinage de 0. Le voisinage de la gaussienne (on parle de l'intervalle $[-3, 3]$) est assez large, tandis que le voisinage de g_α est plus concentré autour de 0 (on constate un pic en 0 de la fonction g_α). Donc bien que la queue de distribution gaussienne soit la plus faible, le comportement autour de 0, handicape cette loi. Nous avons aussi fait apparaître la loi de Cauchy qui a le même comportement que la gaussienne au voisinage de 0, mais une queue de distribution lourde.

Dans la figure 1.5, nous avons choisi les paramètres de chaque densité, pour que toutes soient égales à 1 au point 0. Ceci est utilisé comme critère de comparaison de la dispersion (ou la variance, mais n'existe pas pour la loi de Cauchy) de chaque densité.



(a) Régression Gaussienne.



(b) Régression α.

FIGURE 1.4 – Comparaison entre des observations gaussiennes et des observations issues de la régression α pour $\alpha = 1/4$.

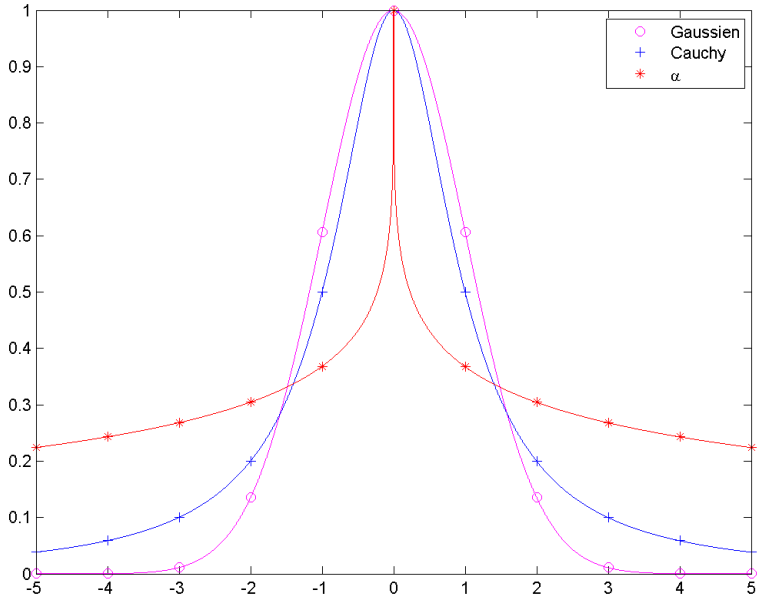


FIGURE 1.5 – Densités gaussienne, de Cauchy et α, $\alpha = 1/4$

1.2.7 Régression Multiplicative Uniforme

Soit l'expérience statistique générée par les observations $\mathcal{Z}_n = (X_i, Y_i)_{i=1,\dots n}$, $n \in \mathbb{N}^*$, où (X_i, Y_i) satisfont à l'équation

$$Y_i = f(X_i) \times U_i, \quad i = 1, \dots, n, \tag{1.2.5}$$

où $f \in \mathbb{H}_d(\beta, L, M, A)$. Les variables aléatoires $(U_i)_{i\in 1,\dots,n}$ sont supposées indépendantes et uniformément distribuées sur $[0, 1]$. Le design X_i est choisi comme dans (1.2.3). Ce modèle est étudié dans le chapitre 4 pour lequel on utilise l'estimateur *bayésien* pour estimer la fonction de régression. Ce type de modèle peut être utilisé dans le cadre des modèles de frontière développés par Simar et Wilson [2000]. Outre le cadre pratique, il est intéressant de voir que les estimateurs linéaires sont inefficaces dans ce modèle (voir Section 2).

Nous verrons que pour ce modèle, la vitesse de convergence minimax est très rapide. Elle est meilleure que toutes celles des modèles précédents. Visuellement, on peut le constater à l'aide de la figure 1.6. En effet, toutes les observations sont en dessous de la fonction de régression f. Et on voit distinctement apparaître la fonction cible, en regardant le maximum par morceaux des observations. L'estimateur bayésien sera, dans ce modèle, beaucoup plus rapide que les estimateurs linéaires (Voir Sections 2.1 et 3.4.4).

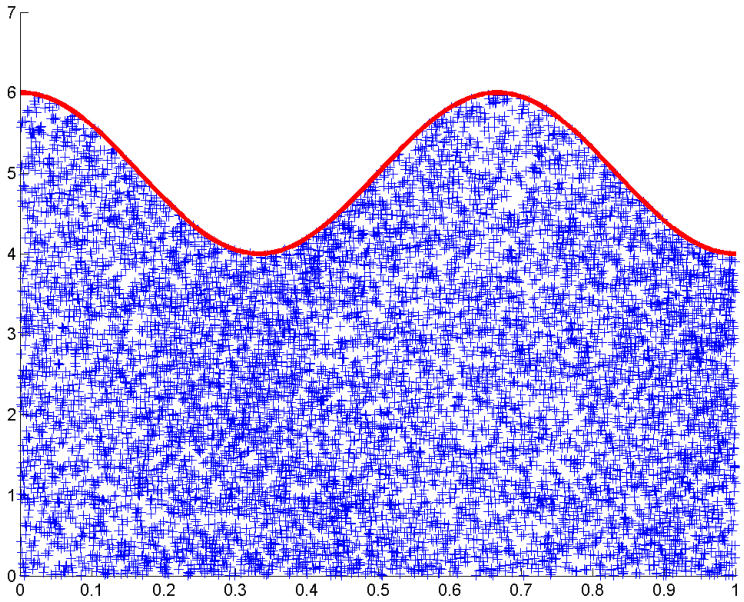


FIGURE 1.6 – Observations d'une fonction bruitée avec un bruit multiplicatif uniforme.

Autres Modèles. On peut citer d'autres modèles des statistiques non-paramétriques : le modèle du bruit blanc et les problèmes inverses. Le modèle de bruit blanc gaussien est défini par l'équation différentielle stochastique :

$$dY_t = f(t)dt + \frac{\sigma}{\sqrt{n}} dW_t, \quad t \in \mathbb{R}^d,$$

où f est la fonction à estimer à partir des observations $(Y_t)_{t\in\mathbb{R}^d}$, W_t est un champ Brownien sur $\mathbb{R}^d$ (Processus de Wiener), $\sigma > 0$ et $n \in \mathbb{N}^*$. Le modèle de bruit blanc gaussien joue un rôle important en Statistiques (voir Has'minskii et Ibragimov [1981]). L'avantage de ce modèle est qu'il est simple à utiliser et qu'il approche d'autres modèles statistiques, en particulier la régression *gaussienne* (Brown et Low [1996] et Nussbaum [1996]).

Pour ce modèle, les estimateurs à noyau sont utilisés pour l'adaptation (voir les travaux les plus récents de Kerkyacharian, Lepski, et Picard [2001], Klutchnikoff [2005], Goldenshluger et Lepski [2008], Goldenshluger et Lepski [2009a]).

Les problèmes inverses sont le résultat de l'équation suivante :

$$dY_t = Af(t)dt + \frac{\sigma}{\sqrt{n}} dW_t, \quad t \in \mathbb{R}^d,$$

où A est un opérateur linéaire défini sur un espace fonctionnel. L'heuristique de ce modèle est que l'on observe indirectement la fonction cible, à travers une transformation de celle-ci, par l'opérateur A connu mais pas forcément inversible. Les problèmes inverses sont très connus pour leurs applications en pratique (équation de la chaleur, tomographie aux rayons X, etc.). Depuis ces dix dernières années, de nombreuses méthodes ont été développées pour l'estimation adaptative de la fonction f (voir par exemple Cavalier et Tsybakov [2002] et Cavalier, Golubev, Picard, et Tsybakov [2002]). Par exemple, la méthode de *Stein par blocs* est fréquemment utilisée. Le modèle est projeté dans une base et on obtient le modèle de suite gaussienne. On observe alors les coefficients de la fonction f, multipliés par les valeurs singulières de l'opérateur A, auxquels on ajoute un bruit gaussien. Une autre méthode, développée pour les problèmes inverses très mal posés (l'opérateur A n'est pas inversible), est appelée *enveloppe du risque* et introduite par Cavalier et Golubev [2006], Cavalier [2008] et Marteau [2009].

1.3 Approche Localement Paramétrique

Dans cette section, nous présentons l'approche *localement paramétrique* (en anglais : *local parametric fitting*). Plus précisemment, nous traitons l'approche *polynomiale locale* qui nous permet d'approximer localement la fonction f par un polynôme noté f_θ lui-même inconnu. Ensuite, nous proposons deux estimateurs permettant d'estimer les coefficients du polynôme d'approximation f_θ. Cette technique est appelée *Méthode des polynômes locaux*, elle a été utilisée pour la première fois par Katkovnik [1985] (pour plus de détails voir Tsybakov [2008]). Notons que, en pratique, le degré du polynôme d'approximation est choisi assez petit (degré $= 0, 1$ ou encore 2), ceci est dû au temps de calculs qui est exponentiellement grand par rapport au nombre de coefficients à estimer.

L'approche localement paramétrique ne s'arrête pas qu'aux polynômes. On peut imaginer approximer la fonction f par tout autre objet paramétrique. Par exemple, on peut décomposer f dans une base orthogonale où l'on ne gardera qu'un nombre fini de coefficients.

Méthode des Polynômes Locaux. On se place dans le modèle de régression générale défini dans (1.2.1). Soit $f \in \mathbb{H}_d(\beta, L, M)$ avec $L, M > 0$ et $\beta \in [0, b[$ où $b \in \mathbb{N}^*$ peut être choisi arbitrairement grand (b est la régularité maximale des fonctions que l'on étudie). Soit

$$V_h(y) = \left\{ X_i \in \bigotimes_{j=1}^{d} \big[y_j - h/2, y_j + h/2\big] \cap [0,1]^d \right\}, \tag{1.3.1}$$

un voisinage de y tel que $V_h(y) \subseteq [0,1]^d$, où est $h \in (0,1)$ un scalaire donné. Soit

$$D_b = \sum_{m=0}^{b} \binom{m+d-1}{d-1}. \tag{1.3.2}$$

Soit $K(z)$, $\forall z \in \mathbb{R}^d$ le vecteur de dimension D_b des polynômes du type suivant :

$$K^\top(z) = \left(\prod_{j=1}^{d} z_j^{p_j}, \;\; (p_1, \dots, p_d) \in \mathbb{N}^d : \; 0 \le p_1 + \cdots + p_d \le b \right),$$

où le signe $\top$ représente la fonction tranposition. Ensuite, pour tout $t \in \mathbb{R}^{D_b}$, où $t^\top = \big(t_{p_1,\dots,p_d}, \;\; (p_1, \dots, p_d) \in \mathbb{N}^d : \; 0 \le p_1 + \cdots + p_d \le b\big)$, on définit le polynôme local

$$f_t(x) \;\; = \;\; t^\top K\left(\frac{x-y}{h}\right) \mathbb{I}_{V_h(y)}(x), \quad x \in [0,1]^d, \tag{1.3.3}$$

où $\mathbb{I}$ est la fonction indicatrice. Notons que $f_t(y) = t_{0,\dots,0}$. Introduisons l'ensemble des coefficients

$$\Theta\big(M\big) = \big\{ t \in \mathbb{R}^{D_b} : \;\; \|t\|_1 \le M \big\} \subset \mathbb{R}^{D_b}, \tag{1.3.4}$$

où $\|.\|_1$ est la norme ℓ_1 sur $\mathbb{R}^{D_b}$. Remarquons que pour tout $f \in \mathbb{H}_d(\beta, L, M)$

$$\exists \theta = \theta(f, y, h) \in \Theta\big(M\big) : \quad \sup_{x \in V_h(y)} \big|f(x) - f_\theta(x)\big| \leq Ldh^\beta.$$

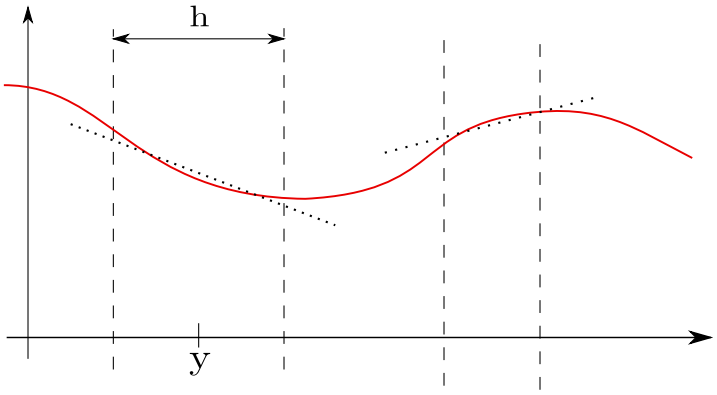


FIGURE 1.7 – Illustration d'une approximation locale d'une fonction (ligne) par une droite ou un polynôme d'ordre 1 (pointillé)).

Le polynôme local f_θ peut être vu comme une approximation localement paramétrique de la fonction de régression f sur le voisinage $V_y(h)$. Par exemple, on peut choisir pour f_θ le polynôme de Taylor de la fonction f au point y (voir Figure 1.7).

Comme nous l'avons précisé en début de chapitre, la restriction aux polynômes locaux n'est pas nécessaire, si l'on trouve un autre objet paramétrique capable d'approcher la fonction cible (par exemple, une décomposition en base d'ondelettes à coefficients finis).

L'idée principale est que si h est choisi suffisament petit, les observations originales (1.2.1) sont bien approximées par le modèle paramétrique $\mathcal{Y}_i$ de densité $g(., f_\theta(X_i))$ dans lequel, sous certaines conditions sur $g(.)$, les estimateurs de *Huber* et *bayésien* sont optimaux au sens des vitesses de convergences. Dans la suite nous présentons les estimateurs qui nous permettent d'estimer θ, en rappelant pour commencer, la définition d'un estimateur.

Définition 3. *On dit que $\tilde{f}$ est un estimateur si $\tilde{f}(.) = \tilde{f}\big(., \mathcal{Z}_n\big)$ est une fonction mesurable des observations.*

Définition 4. *On dit que $\tilde{f}$ est un estimateur linéaire si il existe $K : [0,1]^d \times [0,1]^d \to \mathbb{R}$ telle que :*

$$\tilde{f}(y) = \sum_{i=1}^{d} K\big(X_i, y\big)\, Y_i, \quad \forall f \in \mathbb{H}_d(\beta, L), \quad \forall y \in [0,1]^d.$$

On peut trouver cette définition dans les notes de Nemirovski [2000].

Définition 5. *On appelle $f_{\hat\theta}$ estimateur par polynômes locaux si $\hat\theta$ est solution du problème de minimisation suivant :*

$$\min_{t\in\Theta(M)} \ell\big(\mathcal{Z}_n, f_t\big),$$

où $\ell(.,.)$ est un critère à choisir.

Dans cette thèse, on rencontre les critères bayésien, des moindres carrés, des valeurs absolues et de Huber (voir ci-dessous).

1.3.1 Estimateur Bayésien

Considérons, dans cette section, le modèle de régression *générale* définie dans la section 1.2.2 et soit $\mathbb{E}_f = \mathbb{E}_f^n$ l'espérance mathématique par rapport à la loi de probabilité $\mathbb{P}_f = \mathbb{P}_f^n$ des observations $\mathcal{Z}_n$. Définissons *la pseudo vraisemblance*

$$L_h\left(t, \mathcal{Z}_n\right) = \prod_{X_i\in V_h(y)} g\big(Y_i, f_t(X_i)\big), \quad t\in\Theta\big(M\big), \tag{1.3.5}$$

et le *critère bayésien*

$$\pi_h(t) = \int_{\Theta(M)} \|t-u\|_1 \ \big[L_h\big(u, \mathcal{Z}_n\big)\big]^{1/m} du, \quad t\in\Theta\big(M\big). \tag{1.3.6}$$

où m est une constante positive choisie dans la suite (voir Hypothèses 1, Chapitre 3).

Soit $\hat\theta(h)$ la solution du problème de minimisation suivant :

$$\hat\theta(h) = \arg\min_{t\in\Theta(M)} \pi_h(t), \tag{1.3.7}$$

où $\|.\|_1$ est la norme ℓ_1 sur $\mathbb{R}^{D_b}$. L'*estimateur localement bayésien* $\hat f^h(y)$ de $f(y)$ est défini maintenant comme $\hat f^h(y) = \hat\theta_{0,\dots,0}(h)$.

Nous constatons que la vraisemblance joue le rôle de la loi *a priori* sur le paramètre à estimer. Il est possible de rajouter un autre *a priori* sur le paramètre. En effet, on a l'habitude de voir (Has'minskii et Ibragimov [1981]) un critère de la forme $\int \|t-u\|_1 L_h(u)q(u)du$ où $q(.)$ est une densité sur $\Theta(M)$ (dans notre cas $q(.)\equiv 1$). Dans l'approche bayésienne, le but est de choisir de façon optimale cette densité. Ici, nous nous intéressons seulement au problème d'optimalité des vitesses de convergence. On remarque aussi que la vraisemblance est élevée à la puissance $1/m$ (m est à choisir). C'est une nouvelle façon de considérer l'estimateur de type *bayésien*, traditionnellement $m=1$. Le choix de m permet d'affaiblir les hypothèses sur la densité $g(.)$ et d'avoir une meilleure compréhension de la démonstration. Dans le chapitre 4 nous prenons $m=1$, en revanche dans le chapitre 3 nous considérons m à choisir (par exemple $m=2$ pour la régression gaussienne). Le chapitre 4 est un cas particulier du chapitre 3 du point de vue des modèles, mais les démonstrations sont plus complexes dans le chapitre 4 car $m=1$ est mal ajusté.

Le choix de la fonction de perte, ici la norme ℓ_1, n'est pas restrictif. Ceci nous permet d'améliorer les constantes de majoration du risque. On peut prendre notamment la norme ℓ_2 qui permet de donner une forme explicite de $\hat{\theta}$ pour l'estimation paramétrique unidimensionnelle (Has'minskii et Ibragimov [1981]) :

$$\hat{\theta}(h) = \frac{\int_{\Theta(M)} u \, \left[L_h\big(u, \mathcal{Z}_n\big)\right]^{1/m} du}{\int_{\Theta(M)} \left[L_h\big(v, \mathcal{Z}_n\big)\right]^{1/m} dv}.$$

Cet estimateur est aussi connu sous le nom d'*estimateur de Pitman* (voir Has'minskii et Ibragimov [1981]). Cette version de l'estimateur bayésien donne un estimateur linéaire pour la régression gaussienne. Mais cela ne dépend pas de la norme ℓ_2 utilisée ici. En effet la forme de l'estimateur bayésien dépend directement de la densité $g(.,.)$ des observations et non de la fonction de perte utilisée pour sa construction. Notons que cette méthode est similaire à une approche localement paramétrique établie à partir d'estimateurs du maximum de vraisemblance récemment développés par Polzehl et Spokoiny [2006] et Katkovnik et Spokoiny [2008] pour les *modèles statistiques réguliers*.

Le principal avantage de ce nouvel estimateur *bayésien*, est qu'il atteint la vitesse de convergence optimale du modèle. Par exemple, pour les régressions *gaussienne* et *multiplicative uniforme*, notre estimateur atteint respectivement les vitesses minimax (voir Définition 8) $n^{-\frac{\beta}{2\beta+d}}$ et $n^{-\frac{\beta}{\beta+d}}$ sous les hypothèses 3. Nous pouvons dire que l'estimateur *bayésien* concurrence (fait mieux ou aussi bien que) les estimateurs linéaires dans certains modèles (voir Section 2.1).

Estimateur du Maximum de Vraisemblance

Il est très important de se poser la question suivante : existe-t-il d'autres estimateurs optimaux comme l'estimateur bayésien ?

La réponse à cette question est positive. En effet, l'exemple le plus connu est l'estimateur du maximum de vraisemblance. Cet estimateur concurrence les estimateurs linéaires au même titre que l'estimateur bayésien. On peut expliquer ce phénomène par le fait que ces deux estimateurs utilisent l'information contenue dans la vraisemblance du modèle et donc sont capables d'améliorer la vitesse. Définissons cet estimateur

$$\mathring{f}^h(y) = \arg \max_{t \in \Theta(M)} L_h\left(t, \mathcal{Z}_n\right)$$

Dans le cas paramétrique, si l'on considère le modèle $Y_i \sim \mathcal{U}_{[0,\theta]}$, $i = 1, ..., n$, l'estimateur du maximum de vraisemblance correspondant est $\max_{i=1,...,n} Y_i$. Comme cet estimateur est basé sur le maximum des observations, donc non-linéaire, il est possible d'améliorer la vitesse de convergence. Ainsi, l'estimateur $\max_{i=1,...,n} Y_i$ converge vers θ à la vitesse $1/n$ contre $1/\sqrt{n}$ pour les estimateurs linéaires.

On peut généraliser cette approche au cas non-paramétrique en utilisant comme estimateur $\max_{X_i \in V_h(y)} Y_i$ (estimateur localement constant). Il semble très difficile de pouvoir généraliser cette approche pour un estimateur localement polynômial d'ordre D_b quelconque.

Un autre problème est qu'il faut supposer une hypothèse supplémentaire (par rapport à l'estimateur bayésien) de continuité sur la vraisemblance. Has'minskii et Ibragimov [1981] (voir Chapitre 1, Section 5, Théorème 5.1) supposent cette hypothèse pour l'estimation paramétrique. Mais jusqu'à présent, on ne trouve pas dans la littérature une extension de l'estimateur du maximum de vraisemblance local comme nous l'avons fait pour l'estimateur bayésien dans le problème non-paramétrique.

Polzehl et Spokoiny [2006] utilisent l'estimateur du maximum de vraisemblance dans les *modèles de famille exponentielles* dans le cadre localement paramétrique (fonctions localement constantes). Les auteurs donnent des résultats très généraux sur ces modèles en supposant l'existence de *l'information de Fisher* (*modèle régulier*). Mais les vitesses de convergences ne sont pas améliorées et restent les mêmes que pour les estimateurs linéaires.

Ainsi, pour la régression multiplicative uniforme non-paramétrique, il n'est pas prouvé que l'on peut utiliser l'estimateur du maximum de vraisemblance et cela semble un objectif difficile.

Comme on vient de le voir, les raisons de l'utilisation de l'estimateur bayésien (préféré à l'estimateur du maximum de vraisemblance) sont essentiellement techniques.

1.3.2 Estimateur de Huber

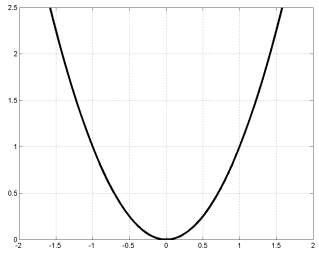
On se place, dans cette section, dans le cadre de la régression additive avec la densité g_ξ inconnue definie dans (1.2.2), Nous supposons vraies les hypothèses 1 sur la densité g_ξ. Nous introduisons une variante de l'estimateur de la médiane, plus communément appelé *estimateur de Huber*, développé pour l'estimation non-paramétrique par Tsybakov [1982a, 1982b, 1983, 1986] et Härdle et Tsybakov [1988, 1992]. Ce dernier donne une règle de sélection locale pour la fenêtre basée sur la méthode *Plug-in* avec des résultats en normalité asymptotique. Un peu plus tôt Hall et Jones [1990] propose d'utiliser la *validation croisée* pour le risque L_2. Plus récemment, Reiss, Rozenholc, et Cuenod [2009] ont développé l'adaptation d'estimateurs robustes, en particulier le critère de Huber mais seulement pour les estimateurs localement constants. Considérons la *fonction de Huber*,

$$Q(z) = \frac{z^2}{2}\,\mathbb{I}_{|z|\le 1} + \left(|z| - \frac{1}{2}\right)\,\mathbb{I}_{|z|>1}, \tag{1.3.8}$$

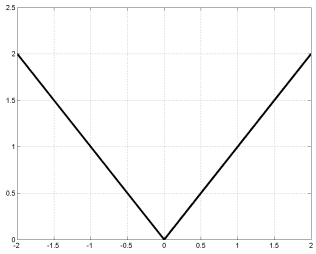
où $\mathbb{I}$ est la fonction indicatrice.

Définissons le *critère de Huber*

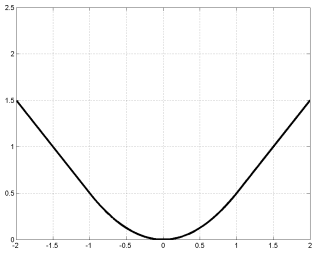
$$\tilde{m}_h(t) = \tilde{m}_h\big(t, \mathcal{Z}_n\big) := \frac{1}{nh^d}\sum_{i=1}^{n} Q\left(Y_i - f_t(X_i)\right)\,\mathbb{I}_{\{X_i \in V_h(y)\}} \tag{1.3.9}$$



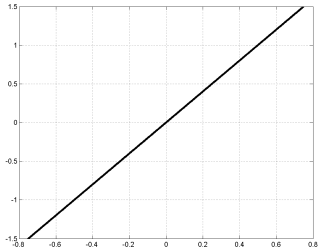
Fonction Carrée z^2.



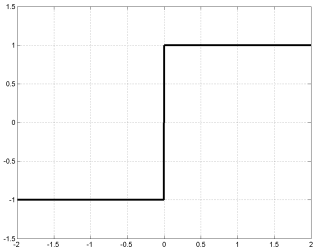
Fonction Valeur Absolue $|z|$.



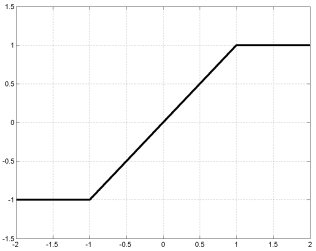
Fonction de Huber Q(z).



Dérivée Carrée $2z$.



Dérivée Valeur Absolue Sign(z).



Dérivée de Huber $Q'(z)$.

FIGURE 1.8 – Courbes des fonctions carré, valeur absolue, de Huber et leurs dérivées.

On introduit le critère de Huber qui allie les propriétés pratiques (robustesse) du critère ℓ_1 et les propriétés théoriques du critère ℓ_2 (dérivabilité). On parle d'estimateur *robuste* si l'estimateur en question n'est pas sensible aux valeurs extrêmes, (voir Rousseeuw et Leroy [1987] et Huber et Ronchetti [2009]).

Ce critère fait partie de la famille des critères par *polynômes locaux*. Par exemple, on peut utiliser le critère ℓ_2 appelée *moindres carrés* ($Q(z) = z^2$, voir Tsybakov [2008], Chapitre 1), ou bien le critère ℓ_1 appelé *valeurs absolues* ($Q(z) = |z|$).

Le critère ℓ_2, qui donne un estimateur fondé sur la moyenne, est beaucoup plus sensible aux valeurs extrêmes (voir Figure 1.8, la dérivée de la fonction carrée). En pratique, quand on cherche un estimateur robuste, on utilise le critère ℓ_1 qui conduit à une estimateur fondé sur la médiane. Mais en estimation non-paramétrique, cela pose un problème car la fonction valeur absolue n'est pas dérivable en 0 (voir Figure 1.8). L'avantage de la fonction de Huber est que sa dérivée est continue sur $\mathbb{R}$, tandis que la dérivée de la fonction valeur absolue n'est pas continue en 0 (voir Figure 1.8). On peut voir dans le chapitre 5, que le contrôle des grandes déviations (Proposition 7) est démontré à l'aide d'un argument de *chaînage*, celui-ci nécessite la continuité de la dérivée de la fonction Q utilisée dans le critère.

Soit $\breve{\theta}(h)$ la solution du problème de minimisation suivant :

$$\breve{\theta}(h) = \arg \min_{t \in \Theta(M)} \tilde{m}_h(t). \tag{1.3.10}$$

L'*estimateur de Huber local* $\breve{f}^h(y)$ de $f(y)$ est défini comme $\breve{f}^h(y) = \breve{\theta}_{0,\dots,0}(h)$.

La fonction de Huber a été introduite par Huber [1964,1981] dans l'idée de développer des estimateurs robustes par rapport aux valeurs extrêmes. Cette idée a été reprise en pratique par Petrus [1999] pour la régression gaussienne dans le cadre de la reconstruction d'images, et par Chang et Guo [2005] dans le modèle paramétrique gaussien pour l'estimation du positionnement par satellite. Nous proposons dans le chaptire 5 de démontrer les performances statistiques de cet estimateur. En outre, nous démontrons que la vitesse de convergence atteinte est $n^{-\frac{\beta}{2\beta+1}}$ pour toute densité g_ξ vérifiant les hypothèses 1. Tsybakov [1982a] a démontré que cette vitesse est minimax pour le risque en probabilité avec la condition suivante sur g_ξ.

$$\exists p_* > 0,\ v_0 > 0\ :\quad \int_{\mathbb{R}} g_\xi(u) \ln \frac{g_\xi(u)}{g_\xi(u+v)} du \leq p_* v^2,$$

pour tout $|v| \leq v_0$. Cette condition n'est autre qu'un contrôle de la *distance de Kullback* entre les deux probabilités (Voir Tsybakov [2008], Chapitre 2). On peut montrer que les densités gaussienne et de cauchy vérifient cette condition. Mais par exemple, pour la régression α ($\alpha < 1/2$), cette condition n'est pas vérifiée.

Un cadre plus général est développé par Arias-Castro et Donoho [2009] pour les médianes locales itérées, justement plébiscitée pour leur capacité à traiter les sauts, et pas uniquement les zones homogènes. Leurs résultats sont aussi de type minimax pour des fonctions localement Lipschitz (dans le cadre unidimensionnel et bidimensionnel). Arias-Castro et Donoho [2009] montrent que le risque est du même ordre de grandeur pour les méthodes par moyennes locales ou par médianes locales dans le cadre régulier (fonctions globalement lipschitziennes). En revanche, si la fonction cible est discontinue et si les courbes régulières sont suffisamment séparées, ils montrent qu'une double itération de la méthode par médianes locales permet de diminuer la vitesse du risque par rapport au cas par moyennes ou médianes locales. Ce dernier article motive le fait que le critère ℓ_1 est plus performant que le critère ℓ_2, notamment quand celui-ci échoue (par exemple, fonctions discontinues ou bruit sans moment).

Notons que la fonction de *Huber* utilise deux régimes, l'un reposant sur le critère ℓ_1 et l'autre sur le cirtère ℓ_2. En effet, on constate que le critère ℓ_2 traite les petits résidus (sur l'intervalle $[-1,1]$) et le critère ℓ_1 les résidus plus importants notamment les valeurs extrêmes. La norme ℓ_1 permet d'obtenir un estimateur plus robuste (que les estimateurs linéaires par exemple). L'utilisation du critère ℓ_2 au voisinage de 0 implique que la dérivée de la fonction de *Huber* existe et est C^0 en 0. Ce dernier point est nécessaire pour contrôler les grandes déviations (Proposition 7) de l'estimateur de Huber. Mais, on peut utiliser le critère ℓ_2 sur n'importe quel intervalle $[-K,K]$ où K est un paramètre de transition entre les deux normes. En effet, il est facile d'observer que les résultats théoriques sont vérifiés pour toute constante $K > 0$. En pratique le problème du choix de cette constante est un problème ouvert.

L'avantage, de cette approche, est que l'on peut estimer la fonction de régression dans le modèle de *régression additive* avec densité inconnue, i.e. l'estimateur de *Huber* ne dépend pas de la densité g_ξ du bruit.

Dans le chapitre 2 (Section 2.2), nous proposons un estimateur de Huber adaptatif, construit à l'aide de la méthode de Lepski. Cette méthode nécessite un nouveau type d'inégalité de concentration de cet estimateur, développé dans le chapitre 5.

1.4 Mesure de l'Erreur

Après avoir développé les méthodes de construction d'estimateurs, on veut désormais choisir celui qui se sera le plus performant dans un modèle donné. Pour cela, il nous faut introduire des outils mathématiques permettant d'évaluer la performance des estimateurs et de les comparer pour choisir le meilleur. L'approche *minimax* est le premier outil développé (voir Stone [1980], Has'minskii et Ibragimov [1981] et Stone [1982]) dans ce sens pour l'estimation non-paramétrique. Pour l'adaptation, nous parlerons d'*adaptation au sens minimax* (voir Efromovich et Pinsker [1984], Lepski [1990]). Ces deux notions sont présentées dans les sections suivantes et sont utilisées tout au long de cette thèse. Notons qu'un paragraphe est consacré à l'approche *oracle* et aux *inégalités oracle*.

1.4.1 Approche Minimax Ponctuelle

Pour mesurer les performances des procédures d'estimation sur l'espace $\mathbb{H}_d(\beta, L, M)$, on utilise l'approche *minimax*. Fixons les paramètres de l'espace $\beta,\ L,\ M > 0$.

Rappelons que $\mathbb{E}_f = \mathbb{E}_f^n$ est l'espérance mathématique par rapport à la loi de probabilité $\mathbb{P}_f = \mathbb{P}_f^n$ des observations $\mathcal{Z}_n$ satisfaisant le modèle de régression *générale* (1.2.2). Premièrement nous définissons le risque maximal sur $\mathbb{H}_d(\beta, L, M)$ correspondant à l'estimation de la fonction f en un point donné $y \in [0,1]^d$. On parlera d'*estimation ponctuelle*. Deuxièmement, on définit le risque minimax, et enfin les notions, de vitesse de convergence minimax et estimateurs minimax, sont données.
Soit $\tilde{f}$ un estimateur construit à partir des observations $\mathcal{Z}_n$ et choisi de façon arbitraire.

Définition 6. *Soit $q > 0$, on appelle risque maximal de l'estimateur $\tilde{f}$ sur $\mathbb{H}_d(\beta, L, M)$ la quantité suivante*

$$R_{n,q}\big[\tilde{f}, \mathbb{H}_d(\beta, L, M)\big] = \sup_{f \in \mathbb{H}_d(\beta,L,M)} \mathbb{E}_f \big|\tilde{f}(y) - f(y)\big|^q.$$

En général, on s'intéresse à la vitesse de convergence ψ_n vers 0 du risque maximal d'un estimateur $\tilde{f}$, nous cherchons un estimateur qui atteint la "meilleure" vitesse de convergence pour le risque maximal. Remarquons que cette vitesse sera donnée pour la pire fonction à estimer dans la classe de Hölder $\mathbb{H}_d(\beta, L, M)$, cette approche est souvent considérée comme pessimiste, mais elle est nécessaire lorsque l'on suppose que la fonction f appartient à un espace fonctionnel. Il est bon de noter que le risque maximal est utilisé dans cette thèse avec la semi-norme ponctuelle. L'avantage principale de ce risque est de choisir de façon locale la fenêtre d'estimation. Donc on aura un estimateur qui s'adapte localement à la régularité de

la fonction f à estimer. Les résultats de type borne supérieure pour ce risque sont souvent les plus techniques à établir, il est possible d'obtenir le risque "global" en norme L_p en intégrant le risque "ponctuel" et en utilisant le théorème de Fubini.

Définition 7. *On appelle risque minimax sur* $\mathbb{H}_d(\beta, L, M)$ *la quantité suivante*

$$R_{n,q}\big[\mathbb{H}_d(\beta, L, M)\big] = \inf_{\tilde{f}} R_{n,q}\big[\tilde{f}, \mathbb{H}_d(\beta, L, M)\big],$$

où inf *est pris sur l'ensemble des estimateurs (fonctions mesurables des observations).*

Cette quantité peut être décrite de la façon suivante : "on mesure de manière théorique le risque du meilleur estimateur possible pour la pire fonction à estimer dans une boule de Hölder".

Définition 8. *La suite de normalisations* $\psi_n(\beta)$ *est appelée vitesse de convergence minimax et l'estimateur* $\hat{f}$ *est dit minimax (asymptotiquement minimax) si*

$$\begin{aligned} \liminf_{n\to\infty} \psi_n^{-q}(\beta)\ R_{n,q}\big[\hat{f}, \mathbb{H}_d(\beta, L, M)\big] &> 0; \\ \limsup_{n\to\infty} \psi_n^{-q}(\beta)\ R_{n,q}\big[\hat{f}, \mathbb{H}_d(\beta, L, M)\big] &< \infty. \end{aligned}$$

Pour un modèle donné, notre but est de construire l'estimateur minimax (celui qui atteint la vitesse minimax). Techniquement, il nous faut montrer une borne supérieure du risque maximal et une borne inférieure pour le risque minimax. Ces deux bornes doivent être égales à constante près. Pour la borne inférieure, de nombreuses techniques ont été mises en place, on renvoie le lecteur au livre de Tsybakov [2008] pour une présentation de ces méthodes. Plus récemment, dans l'article Chichignoud [2010a] (Chapitre 4, Section 4.2), nous utilisons le critère introduit par Klutchnikoff [2005] pour obtenir la borne inférieure du risque minimax (voir Chapitre 4). Pour les bornes supérieures, la forme de l'estimateur en question et des techniques classiques sont utilisées comme la décomposition biais-variance.

Dans la majorité des cas, les estimateurs minimax dépendent des paramètres de régularité β et L, ($\hat{f} = \hat{f}_{\beta,L}$). En pratique, le problème est que les paramètres de régularité β, L ne sont pas connus par le statisticien. De plus, ces paramètres ne peuvent pas être estimés au préalable à l'inverse de la borne supérieure M. Dans ce sens, l'*adaptation* a été développée dans les années 90 par Efromovich et Pinsker [1984], Lepski [1990], Donoho et Johnstone [1994], Barron, Birgé, et Massart [1999], etc. Ces méthodes permettent de construire des estimateurs qui ne dépendent pas de la régularité de la fonction cible et qui ont une "bonne" performance.

Approche Maxiset. Plus récemment, la notion de *maxiset* a été introduite par Kerkyacharian et Picard [2002] et largement développée par Autin [2006], Autin, Picard, et Rivoirard [2006] et Autin [2008]. Cette approche est complémentaire aux motivations énoncées

dans ce paragraphe. Etant donnés un estimateur $\hat{f}$ et une vitesse de convergence φ_n, l'idée est de chercher l'espace fonctionnel le plus grand possible sur lequel $\hat{f}$ atteint la vitesse φ_n. Cette notion est beaucoup moins pessimiste que le risque minimax puisqu'un estimateur n'est plus défini par "la plus mauvaise" performance sur un espace préalablement fixé, mais par l'ensemble des fonctions pouvant être estimées par $\hat{f}$ à la vitesse φ_n. Cette approche peut être utilisée pour une large gamme de modèles. Cette approche, qui peut être intéressante pour les praticiens, dans le sens où elle exhibe les fonctions "bien estimées" (à vitesse choisie) par un estimateur choisi. L'inconvénient est que deux estimateurs ne sont pas toujours comparables si leurs maxisets ne sont pas emboîtés (par exemple, seuillage par blocs et seuillage par arbre).

1.4.2 Approche Minimax Adaptative

Dans cette section, nous présentons les contraintes de l'adaptation et les résultats attendus. Le but de l'adaptation au sens minimax est de choisir un estimateur $\hat{f}^*$ qui ne dépend pas de β, L et qui atteint la "meilleure" vitesse de convergence pour tout β, L. Nous verrons qu'il est souvent nécessaire de payer un prix pour l'adaptation, c'est-à-dire que l'on peut construire un estimateur qui ne dépend pas de β, L mais qui atteint seulement la vitesse ψ_n^p minimax pénalisée. La question que l'on se pose naturellement est la suivante : Existe-t-il un estimateur $\hat{f}^*$ qui ne dépend pas de β, L et qui atteint la vitesse minimax ψ_n ? Pour les risques L_p la réponse à cette question est positive, voir Efromovich et Pinsker [1984], Lepski [1991], Donoho, Johnstone, Kerkyacharian, et Picard [1995], Lepski, Mammen, et Spokoiny [1997], Goldenshluger et Nemirovski [1997], Juditsky [1997] et Goldenshluger et Lepski [2008]. Pour le risque ponctuel, le prix à payer n'est pas nul, Lepski [1990,1992a,1992b], Brown et Low [1996], Spokoiny [1996], Lepski et Spokoiny [1997], Tsybakov [1998], Klutchnikoff [2005], Chichignoud [2010a,2010b,2010c]. En général, le prix à payer est une puissance de $\ln n$.

Définition 9. *On appelle $\{\psi_n^p(\beta)\}_{\beta,L}$ une famille de normalisations indexée par les paramètres de régularité. On dit que l'estimateur $\hat{f}^*$ est ψ-adaptatif si et seulement si $\forall \beta, L$ on a*

$$\limsup_{n\to\infty} \left(\psi_n^p(\beta)\right)^{-q} R_{n,q}\left[\hat{f}^*, \mathbb{H}_d(\beta, L, M)\right] < \infty.$$

Remarquons que la vitesse de convergence ne dépend pas de L. C'est en effet souvent le cas : la constante de Lipschitz intervient de façon polynômiale dans les constantes de la borne supérieure.

La recherche de la meilleure famille de normalisations (celle qui tend le plus vite vers 0) est une priorité. L'optimalité de ces familles a été étudiée par Lepski [1992a], Lepski [1992b], Lepski et Spokoiny [1997], Tsybakov [1998], Klutchnikoff [2005] et Chichignoud [2010a]. Pour obtenir les bornes supérieures, le recours à des inégalités de concentration exponentielles est nécessaire (voir Section 1.5). On renvoie le lecteur au livre de Massart [2007] sur les inégalités maximales et de concentration. Rappelons que notre estimateur $\hat{f}^*$ est *adaptatif* pour la pire fonction dans la boule de Hölder. Ce point de vue est souvent considéré comme pessimiste.

Approche Oracle. Il est possible de construire des méthodes adaptatives à toutes fonctions dans un espace de Hilbert H. Etant données une méthode $\hat{f}$ et une famille d'estimateurs Λ, une inégalité du type :

$$\mathbb{E}_f\|\hat{f} - f\|^2 \leq (1+\rho_n) \inf_{\tilde{f}\in\Lambda} \mathbb{E}_f\|\tilde{f} - f\|^2 + T_n, \quad \forall f \in H,$$

et $\rho_n \geq 0$ est appelée **inégalité oracle**.

On appelle *oracle* l'estimateur $\hat{f}_0 = \arg\min_{\tilde{f}\in\Lambda} \mathbb{E}_f\|\tilde{f} - f\|^2$. Connaissant la fonction f, $\hat{f}_0$ correspond au meilleur estimateur possible dans la famille Λ. T_n est un terme résiduel à contrôler.

Dans de nombreuses situations, une inégalité oracle est un très bon moyen d'obtenir des résultats d'adaptation au sens minimax, i.e. en utilisant le fait que l'oracle dans la famille est un estimateur minimax (résultat à démontrer au préalable), on peut montrer que l'estimateur $\hat{f}$ est adaptatif optimal (atteint la vitesse minimax et ne dépend pas de la régularité de l'espace considéré). Noter qu'en général ceci est possible seulement dans le cas de l'estimation globale (en norme L_p), les inégalités d'oracle pour la semi-norme ponctuelle n'apparaissent pas dans la littérature, notamment pour leur différence (prix à payer pour l'adaptation) avec la norme L_p. Ce type de résultat est cependant beaucoup plus fort, il est établi sans aucune restriction sur la fonction f. Par ailleurs, une inégalité oracle présente un caractère non-asymptotique dans la mesure où elle permet de comparer les performances d'une méthode à un estimateur "idéal" $\hat{f}_0$, et ce, quelque soit le nombre d'observations.

L'estimation en grande dimension est aussi très étudiée aujourd'hui. En outre, la notion de *parcimonie* (en anglais : *sparsity*) est systématiquement utilisée pour obtenir des résultats du type *inégalités oracle* (voir par exemple, Bunea, Tsybakov, et Wegkamp [2007]).

Dans cette thèse, les résultats présentés sont asymptotiques par souci de présentation. Néanmoins, au vu des preuves, il est possible d'obtenir des résultats non-asymptotiques, c'est-à-dire que notre borne supérieure sera composée de la vitesse ψ_n^p, plus un reste T_n. En général, on choisit n assez grand pour que le reste soit négligeable. Nous verrons dans la section 1.5.2 que la méthode de Lepski n'a nullement un caractère asymptotique.

Pour l'approche *oracle*, rappelons que la norme $\|.\|$ est une norme L_p. Il est souvent difficile d'obtenir des inégalités oracle avec une semi-norme ponctuelle. Seuls les travaux de Plancade [2008], sur l'estimation de la densité du bruit dans le modèle de régression homoscédastique, vont dans ce sens. Nos résultats sont tous donnés avec le risque ponctuel. Nous verrons que la principale difficulté est d'obtenir un contrôle des grandes déviations avec la semi-norme ponctuelle.

1.5 Adaptation

Cette partie contient un panorama sur les méthodes adaptatives développées ces vingt dernières années. Nous utilisons, dans cette thèse, la méthode de Lepski pour l'adaptation.

Une section y est consacrée pour présenter l'idée générale et les conditions nécessaires à l'application de la méthode.

1.5.1 Généralités

Une des premières méthodes *adaptatives* a été développée par Lepski [1990]. Nous la présentons dans la section 1.5.2. On citera, sans plus de détails, une autre approche possible pour l'adaptation, les méthodes dites *d'agrégation d'estimateurs* (voir notamment Nemirovski [2000], Rigollet et Tsybakov [2007] et Lecué et Mendelson [2009]). Dans la suite, nous présentons deux méthodes d'adaptation très populaires : le *seuillage par ondelettes* et la *sélection de modèles*

Seuillage par ondelettes. Suite à la création des ondelettes par Meyer [1992], les premières méthodes adaptatives fondées sur les ondelettes sont apparues sous l'impulsion de Donoho et Johnstone [1994] et Donoho, Johnstone, Kerkyacharian, et Picard [1995]. L'idée est de projeter la fonction de régression dans une base d'ondelettes et d'estimer les coefficients. On procède par un seuillage (déterministe ou aléatoire) des coefficients pour ne retenir que les plus grands coefficients à savoir ceux qui contiennent l'information. D'un point de vue maxiset, les méthodes de seuillage par arbre ou par blocs ont été jugées très performantes par Autin [2004]. En particulier, leur maxiset contient les espaces de Besov.

Sélection de modèles. On souhaite construire une procédure établie à partir des observations $(Y_i)_i$ qui permet de choisir un estimateur parmi la collection $\{\hat{f}_m : m \in \mathcal{M}\}$ où $\hat{f}_m$ est un minimiseur de γ_n sur S_m, $\{S_m : m \in \mathcal{M}\}$ une collection de modèles. L'objectif est de choisir un estimateur dont le risque est aussi proche que possible de celui de l'oracle. Pour faire ce choix, on procède par pénalisation. On choisit l'estimateur adaptatif $\hat{f}_{\hat{m}}$ tel que

$$\hat{m} = \arg\min \left[\gamma_n\big(\hat{f}_m\big) + \mathrm{pen}(m) \right],$$

où $\gamma_n(.)$ est un contraste empirique et $\mathrm{pen}(m)$ est une pénalité à choisir. Les premiers résultats obtenus par critère pénalisé sont dûs à Akaike [1973] pour l'estimation de densité par vraisemblance pénalisée et à Mallows [1973] pour l'estimation de la fonction de régression dans un cadre gaussien. Pour la pénalité $\mathrm{pen}_{\mathrm{Akaike}}(m) = D_m$ où D_m est la dimension de S_m, Akaike a obtenu le critère AIC et, pour la pénalité $\mathrm{pen}_{\mathrm{Mallows}}(m) = 2\sigma D_m$, le critère s'appelle C_p de Mallows. Birgé et Massart [2001] ont montré que ces critères sont asymptotiquement optimaux à condition que la taille de $\mathcal{M}$ ne soit pas trop grande. Dans un cadre non-asymptotique, Birgé et Massart [2001] ont introduit des pénalités plus générales que celle de Mallows

$$pen_{\mathrm{bm}} = C\,\sigma^2 D_m \left(1 + \sqrt{2L_m}\right)^2, \quad m \in \mathcal{M},$$

où $C > 1$ et $\{L_m : m \in \mathcal{M}\}$ une collection de poids positifs. Birgé et Massart ont obtenu un contrôle sur le risque de f quelque soit la taille de la collection de modèles. La construction de leur procédure se base sur des résultats de concentration de la mesure gaussienne. La qualité d'une procédure de sélection de modèle correspond à sa capacité à choisir un estimateur $\tilde{f}$ parmi une collection $\{\hat{f}_m : m \in \mathcal{M}\}$ qui ait un risque faible. Afin d'évaluer cette qualité, on peut considérer les inégalités de type oracle pour valider théoriquement la qualité de l'estimateur $\tilde{f}$. Il nous faut donc établir des inégalités de la forme

$$\mathbb{E}_f l(f, \tilde{f}) \leq K \inf_{m \in \mathcal{M}} \left\{ \inf_{f' \in S_m} l\big(f, f'\big) + \mathrm{pen}(m) \right\} + R, \quad \forall f,$$

où $l(.,.)$ est une fonction de perte, K est une constante universelle et R un terme de reste. En général, la fonction de perte est celle des moindres carrées ou la divergence de Kullback-Leibler. Le principal avantage est qu'elle n'impose aucune restriction sur la fonction à estimer f. En pratique, cette méthode pose la question du choix de la constante C, qui est une constante de régularité.

Plus récemment, les travaux de Cavalier et Golubev [2006] et Marteau [2009] utilisent une méthode dite *d'enveloppe du risque*, fondée sur la sélection de modèles, pour les problèmes inverses très mal-posés.

1.5.2 Choix de la Fenêtre : Méthode de Lepski

Comme nous l'avons vu, les méthodes utilisant une fenêtre fixe ne peuvent rendre compte de la géométrie (ou de régularité) de la fonction. Une amélioration possible est alors de choisir le voisinage de manière locale, en s'appuyant sur les données observées. L'idée de ce type d'approche est de sélectionner le voisinage le plus pertinent possible à partir des observations. Cette idée bien connue sous le nom de méthode dite de *Lepski* (Lepski [1990]) permet de sélectionner la fenêtre de façon adaptative parmi une grille, cette méthode repose sur la comparaison d'estimateurs appartenant à une grille finie. Cette opération est souvent très chère en complexité de calcul.

Chronologie. La première procédure adaptative fût donnée par Efromovich et Pinsker [1984]. Stone [1982] avait posé, sans la résoudre, la question de l'existence d'un estimateur ne dépendant pas de la régularité de la fonction cible. Lepski [1990] donna une nouvelle procédure qui portera plus tard son nom. Il construit un estimateur adaptatif pour choisir la régularité de Hölder dans le modèle du bruit blanc gaussien pour l'estimation ponctuelle. Il démontra qu'il n'existe pas d'estimateur adaptatif efficace. En effet, il est fréquent de payer un prix $\sqrt{\ln \varepsilon^{-1}}$ pour l'adaptation. C'est-à-dire que l'on ne peut pas faire mieux que la vitesse $(\varepsilon sqrt \ln \varepsilon^{-1})^{\frac{2\beta}{2\beta+1}}$.

Dans l'article Lepski [1991], une procédure plus générale est donnée pour n'importe quel type d'estimateurs d'une fonction. On peut trouver dans cet article les conditions suffisantes

pour obtenir un estimateur adaptatif efficace pour le risque L_p. Dans les deux articles Lepski [1992a,1992b], on peut trouver des exemples où il n'existe pas d'estimateurs efficaces. Cela dépend généralement de la fonction de perte utilisée dans le risque (norme L_p, semi-norme ponctuelle ou encore norme infinie). Ces trois articles traitent entre autres le modèle du bruit blanc, d'estimateurs à noyau et d'estimateurs par polynômes locaux.

Plusieurs articles ont ensuite utilisé la méthode de Lepski pour l'adaptation et ont contribué à l'amélioration de celle-ci. Un choix de la fenêtre est proposé dans Lepski, Mammen, et Spokoiny [1997] pour les estimateurs à noyau. Ils obtiennent sur les espaces de Besov les vitesses adaptatives efficaces, qui étaient déjà connues par Donoho, Johnstone, Kerkyacharian, et Picard [1995] avec les ondelettes. A noter que dans cet article, le choix de la fenêtre est fait avec une méthode ponctuelle et les vitesses sont obtenues en norme L_p.

La régression gaussienne à design fixe a fait l'objet d'un article de Goldenshluger et Nemirovski [1997], où l'estimateur des polynômes locaux est utilisé avec le risque ponctuel et L_p sur les espaces de Sobolev. Spokoiny [1998] a étudié les fonctions discontinues et propose une méthode adaptative reposant sur l'idée de Lepski.

Lepski et Levit [1999] et Kerkyacharian, Lepski, et Picard [2001] proposent des méthodes adaptatives pour les fonctions anisotropes. La nouveauté est dans le caractère multidimensionnel de la fenêtre d'estimation, avec l'introduction d'un estimateur artificiel dans la procédure de Lepski. On note la présence d'inégalités oracle dans Kerkyacharian, Lepski, et Picard [2001] avec le risque L_p. Citons les travaux de Klutchnikoff [2005] sur l'estimation ponctuelle de fonctions de Hölder anisotropes. Plus récemment, l'article de Goldenshluger et Lepski [2008] donne une procédure s'appuyant sur l'idée de Lepski pour un large choix d'estimateurs linéaires sans condition sur le modèle (régression gaussienne, bruit blanc gaussien et modèle de densité). Goldenshluger et Lepski [2009a] donnent une procédure qui permet d'obtenir des inégalités oracle sur une très large famille d'estimateurs linéaires.

Il est possible d'utiliser la règle de Lepski pour les méthodes de seuillage des coefficients d'ondelettes. En effet, Autin [2004] a mis en place une procédure dite *procédure hard tree*, qui utilise la règle de Lepski pour choisir les coefficients à seuiller dans l'arbre. Dans le modèle du bruit blanc gaussien, cette méthode permet d'avoir un maxiset plus important que pour les méthodes de seuillage classiques.

La méthode de Lepski a aussi été utilisée pour les problèmes inverses. L'article de Mathé [2006] est une très bonne description de la méthode dans le cas des problèmes inverses déterministes et aléatoires. Plus récemment, on peut citer une nouvelle procédure de test introduite par Spokoiny et Vial [2009] pour les problèmes inverses.

En pratique, on retrouve souvent la règle de Lepski du fait de son implémentation assez simple, malgré le coût important de la complexité. En effet, il est nécessaire de calculer tous les estimateurs dans la famille que nous considérons. Les travaux de Polzehl et Spokoiny [2000,2003] ont considéré la méthode de Lepski pour le débruitage d'images. L'estimateur du maximum de vraisemblance local est utilisé pour l'adaptation dans les modèles dits *réguliers* dans les articles de Polzehl et Spokoiny [2006] et Katkovnik et Spokoiny [2008], notamment pour les modèles poissonniens adaptés aux images. Dans le même temps, Katkovnik [1999] et,

par la suite Astola, Egiazarian, et Katkovnik [2002] et Astola, Egiazarian, Foi, et Katkovnik [2010], donnent une très bonne illustration de comment utiliser la règle de Lepski en pratique. Par exemple, ils introduisent une nouvelle interprétation de cette méthode que Katkovnik [1999] nomme *ICI* (en anglais : *Intersection of Confidence Intervals*) et utilisent des voisinages anisotropes circulaires, ce qui améliore le débruitage de l'image. Cette méthode, explicitée plus loin, permet de réduire considérablement la complexité de la méthode tout en gardant ces bonnes propriétés théoriques.

FIGURE 1.9 – Illustration de l'effet poivre et sel sur une image traitée avec la méthode de Kervrann et Boulanger [2006] (voir par exemple le coin en bas à droite)

Kervrann et Boulanger [2006] ont utilisé la méthode de Lepski pour le choix de la zone de recherche pour les *Moyennes Non Locales* (en anglais : Non-Local Means ou encore NL-Means, voir Buades, Coll, et Morel [2005]). Ces articles mettent en avant un inconvénient, en pratique, de la méthode : l'image obtenue après traitement a une tendance à être dégradée par un bruit de type poivre et sel (Figure 1.9). Ceci est dû à la procédure, qui avec une probabilité faible, s'arrête trop tôt (voir Description). Pour combler cette défaillance, l'utilisation de filtres médians est souvent préférée à celle des filtres linéaires (voir Astola, Egiazarian, Foi, et Katkovnik [2010]).

L'idée de cette section est de comprendre comment on peut utiliser la méthode de Lepski, donner les idées de la preuve, étayées par quelques calculs, et expliquer les conditions

suffisantes à rassembler pour garantir la performance de la procédure.

Description. Comme nous l'avons précisé ci-dessus, l'approche développée par Lepski [1991] est conçue pour l'adaptation au sens minimax. On peut cependant développer une approche oracle (voir Goldenshluger et Lepski [2009a]). Dans cette partie, nous considérons la régression généralisée définie dans la section 1.2.2 et l'estimateur localement paramétrique $\mathring{f}^h(y) = \mathring{\theta}_{0,\dots,0}(h)$ tel que

$$\mathring{\theta}(h) = \arg \min_{t \in \Theta(M)} \delta\big(t, \mathcal{Z}_n\big),$$

où $\delta(.,.)$ est un contraste empirique. Par exemple, $\delta(.,.)$ peut être le critère des moindres carrés ou bien le critère bayésien (1.3.7) ou encore le critère de Huber (1.3.9). Rappelons que f appartient à une boule de Hölder $H_d(\beta, L)$ (voir Définition 1), où β et L sont inconnus pour le statisticien.

Supposons que nous connaissons déjà la vitesse minimax que nous noterons $\psi_{n,\gamma}(\beta) = n^{-\frac{\beta}{\gamma\beta+d}}$, $1 \le \gamma \le 2$. Nous verrons que, dans le modèle de régression gaussienne, la vitesse est $\psi_{n,2}(\beta)$ et dans le modèle de régression multiplicative uniforme la vitesse est $\psi_{n,1}(\beta)$ (voir Section 3.4). Nous supposons aussi connue la construction de l'estimateur minimax noté $\mathring{f}_{h(\beta,L)}$ où $h(\beta, L)$ est l'argument qui permet d'équilibrer le biais et la variance de notre estimateur que l'on peut écrire comme la solution du problème de minimisation suivant :

$$h(\beta, L) = (L^\gamma n)^{-\frac{1}{\gamma\beta+d}} = \arg\min_{h>0} \; Ldh^\beta + \big(nh^d\big)^{-\frac{1}{\gamma}}. \tag{1.5.1}$$

L'estimateur $\mathring{f}_{h(\beta,L)}$ dépend explicitement de β, L à travers la fenêtre $h(\beta, L)$. Choisir la régularité de façon adaptative revient alors à choisir la fenêtre. La procédure que nous présentons est celle développée par Lepski, Mammen, et Spokoiny [1997] que l'on appelle abusivement *méthode de Lepski*. Cette procédure a été donnée spécialement pour le choix de la fenêtre, et donc pour les estimateurs localement paramétriques. Néanmoins, l'idée peut être utilisée dans un autre contexte. Spokoiny et Vial [2009] utilisent le principe de *Lepski* pour les problèmes inverses. Leur estimateur est la somme successive des coefficients bruitées de la fonction cible. Cet estimateur s'arrête au N-ième coefficient. La question est de savoir à quel rang s'arrêter. Les auteurs développent une méthode pour tester si l'on peut passer au rang supérieur ou non. Ceci est fait à partir de l'idée de Lepski [1990].

L'idée de l'adaptation est la suivante : on dispose d'une famille d'estimateurs $\left\{\mathring{f}^h\right\}_{h\in\mathcal{H}_n}$, où $\mathcal{H}_n \subseteq [0,1]$ tel que $h(\beta, L) \in \mathcal{H}_n$. Cette famille contient un oracle, ici l'estimateur minimax $\mathring{f}_{h(\beta,L)}$. Nous supposons inconnues les quantités β et L, et on admet connaître l'information supplémentaire que $\beta \in]0, b]$ et $L > 0$ où $b \in \mathbb{N}^*$ est connu. Pour la contante L, aucune restriction n'est imposée, en effet, son rôle est minimisé dans le choix de la fenêtre, du fait qu'elle n'intervient pas dans la vitesse de convergence.

Notons $\mathcal{H}_n = [h_{\min}, h_{\max}]$ où $h_{\min} < h_{\max}$ sont à choisir pour que $h(\beta, L) \in \mathcal{H}_n$. La procédure de Lepski commence par la construction d'une grille sur l'ensemble $\mathcal{H}_n$ des

fenêtres, soit

$$h_k = 2^{-k} h_{\max}, \quad k = 0, \dots, \mathbf{k}_n, \tag{1.5.2}$$

où $\mathbf{k}_n$ est le plus grand entier tel que $h_{\mathbf{k}_n} \in \mathcal{H}_n$. Soit

$$\mathcal{F} = \left\{ \mathring{f}^{(k)}(y) = \mathring{f}^{h_k}(y), \quad k = 0, \dots, \mathbf{k}_n \right\}, \tag{1.5.3}$$

la grille construit sur notre famille d'estimateurs localement paramétriques.

Objectif. Le but de l'adaptation est de choisir un estimateur $\mathring{f}^{(\hat{k})} \in \mathcal{F}$ où $\hat{k}$ est une fonction mesurable des observations $\mathcal{Z}_n$. Une fois la procédure établie, il nous faut vérifier que notre estimateur est bien *adaptatif* (voir Définition 9). C'est-à-dire qu'il atteint simultanément sur chaque espace $\mathbb{H}_d(\beta, L)$ la vitesse $\psi_n^p(\beta)$, $\forall \beta \in [0, b[$. Notons que cette famille de normalisations doit être optimale (voir par exemple Chapitre 4). On peut écrire la vitesse adaptative de la façon suivante :

$$\psi_n^p(\beta) = \left(n^{-1} \rho_n(\beta)\right)^{\frac{\beta}{\gamma\beta + d}},$$

où $\rho_n(\beta)$ est un terme dit de *pénalité* ou de *prix à payer* pour l'adaptation. En général, pour l'estimation ponctuelle, on choisit la pénalité égale à $\left[1 + (b - \beta) \ln n\right]^{1/\gamma}$ (voir Lepski [1990,1992a,1992b], Brown et Low [1996], Spokoiny [1996], Lepski et Spokoiny [1997], Tsybakov [1998], Klutchnikoff [2005], Chichignoud [2010a,2010b,2010c]), la description de la procédure permettra d'expliquer ce choix. En norme L_p, la pénalité est souvent égale à 1 (voir Efromovich et Pinsker [1984], Lepski [1991], Donoho, Johnstone, Kerkyacharian, et Picard [1995], Lepski, Mammen, et Spokoiny [1997], Goldenshluger et Nemirovski [1997], Juditsky [1997] et Goldenshluger et Lepski [2008]).

Règle de sélection pour deux fenêtres.

Pour commencer considérons la méthode dans le cas simple du choix entre deux estimateurs. Notons les $\mathring{f}^{(1)} = \mathring{f}^{h_1}$, $0 < h_1 < 1$ et $\mathring{f}^{(2)} = \mathring{f}^{h_2}$, $h_1 < h_2 < 1$, on est dans le cas où $\mathrm{Card}(\mathcal{F}) = 2$. On dispose de l'information supplémentaire : "un des deux estimateurs est minimax sur $\mathbb{H}_d(\beta, L)$" par construction de $\mathcal{F}$. Nous allons construire une procédure pour choisir un estimateur parmi les deux pour atteindre la vitesse adaptative. Soit la règle de sélection suivante :

$$f^* = \begin{cases} \mathring{f}^{(2)}, & \left|\mathring{f}^{(1)}(y) - \mathring{f}^{(2)}(y)\right| \le C\, \sigma_p(h_1) \\ \mathring{f}^{(1)}, & \left|\mathring{f}^{(1)}(y) - \mathring{f}^{(2)}(y)\right| > C\, \sigma_p(h_1) \end{cases}$$

Nous appelons *écart-type* $\sigma_h = \left(nh^d\right)^{-1/\gamma}$ et *écart-type pénalisé* $\sigma_p(h) = \mathrm{pen}_h \times \sigma(h)$ avec pen_h une pénalisation à choisir en fonction du problème. C est une constante que l'on peut voir comme un paramètre de seuillage de la méthode. Nous choisirons celle-ci plus tard. Nous pouvons garantir la performance de notre estimateur par le résultat suivant.

Proposition 1. *Pour tout $\beta, L \in [0, b[\times\mathbb{R}_+$ on a*

$$\sup_{f\in\mathbb{H}_d(\beta,L)} \mathbb{E}_f\big|f^*(y) - f(y)\big| \leq (2c_1 + c_2)\psi_n^p(\beta) + R_n,$$

où c_1, c_2 sont des constantes positives et R_n est un terme de reste que l'on définira dans la suite.

Avec ce proposition, l'estimateur f^* atteint la vitesse $\psi_n^p(\beta)$ pour tout $\beta, L \in [0, b[\times\mathbb{R}_+$.

Idée de la preuve. Soit $h^p(\beta, L) = \left(\frac{\rho_n(\beta)}{L^\gamma n}\right)^{-\frac{1}{\gamma\beta+d}}$ la fenêtre minimax pénalisée. Dans ce paragraphe, nous expliquons comment fonctionnne la méthode de Lepski. On peut comprendre comment choisir la vitesse pénalisée, la constante C et la variance pénalisée.

1er cas : $\mathring{f}^{(1)}$ est minimax pénalisé, c'est-à-dire que $h_1 = h^p(\beta, L)$. Définissons l'événement

$$A = \left\{\big|\mathring{f}^{(1)}(y) - \mathring{f}^{(2)}(y)\big| \leq C\,\sigma_p(h_1)\right\}.$$

1) Si A^c est réalisé alors la règle de sélection donne $f^* = \mathring{f}^{(1)}$ (la règle a choisi le bon estimateur). Comme $\mathring{f}^{(1)}$ est minimax, on a

$$\mathbb{E}_f\big|f^*(y) - f(y)\big|\mathbb{I}_{A^c} \leq \mathbb{E}_f\big|\mathring{f}^{(1)}(y) - f(y)\big| \leq c_1\psi_n^p(\beta).$$

2) Si A est réalisé alors $f^* = \mathring{f}^{(2)}$ (ici la procédure s'est trompée d'estimateur), nous allons regarder quelle est la perte dûe à cette erreur.

$$\mathbb{E}_f\big|f^*(y) - f(y)\big|\mathbb{I}_A = \mathbb{E}_f\big|\mathring{f}^{(2)}(y) - f(y)\big|\mathbb{I}_A \leq \mathbb{E}_f\big|\mathring{f}^{(2)}(y) - \mathring{f}^{(1)}(y)\big|\mathbb{I}_A + \mathbb{E}_f\big|\mathring{f}^{(1)}(y) - f(y)\big|\mathbb{I}_A.$$

Sous l'événement A, le premier terme est contrôlé par $C\,\sigma_p(h_1)$ et comme $\mathring{f}^{(1)}$ est minimax, il existe une constante c_1 telle que

$$\mathbb{E}_f\big|f^*(y) - f(y)\big|\mathbb{I}_A \leq C\,\sigma_p(h_1) + c_1\psi_n^p(\beta).$$

Nous pouvons montrer que si $\mathrm{pen}_{h^p(\beta,L)} = \rho_n(\beta)$ alors il existe une constante c_2 telle que

$$\sigma_p(h_1) = \sigma_p(h^p\big(\beta, L)\big) \leq c_2\psi_n^p(\beta),$$

ainsi on obtient avec les deux inégalités ci-dessus.

$$\mathbb{E}_f\big|f^*(y) - f(y)\big|\mathbb{I}_A \leq (c_1 + c_2)\,\psi_n^p(\beta).$$

2ème cas : $\mathring{f}^{(2)}$ est minimax pénalisé ($h_2 = h^p(\beta, L)$). Dans ce cas, la probabilité que l'événement A^c se réalise est faible.

1) Si A est réalisé alors $f^* = \mathring{f}^{(2)}$ (la procédure donne le bon estimateur). Ainsi

$$\mathbb{E}_f\big|f^*(y) - f(y)\big|\mathbb{I}_A \le \mathbb{E}_f\big|\mathring{f}^{(2)}(y) - f(y)\big| \le c_1\psi_n^p(\beta).$$

2) Si A^c est réalisé alors la règle de sélection donne $f^* = \mathring{f}^{(1)}$ (la procédure s'est trompée). En utilisant l'inégalité de Cauchy-Schwarz, on obtient :

$$\mathbb{E}_f\big|f^*(y) - f(y)\big|\mathbb{I}_{A^c} \le \left(\mathbb{E}_f\big|\mathring{f}^{(1)}(y) - f(y)\big|^2\right)^{1/2}\sqrt{\mathbb{P}_f\{A^c\}}.$$

Le premier facteur (risque de l'estimateur $\mathring{f}^{(1)}$) tend vers 0 (car c'est un estimateur de la fonction f avec une mauvaise fenêtre) mais pas à la bonne vitesse. Il est facile de montrer que ce facteur est néanmoins borné par une constante K. Dans certains exemples, on peut le calculer explicitement et connaître sa vitesse. Le second terme $\sqrt{\mathbb{P}_f\{A^c\}}$ va compenser la perte. Nous pouvons donner une condition suffisante pour que la règle de sélection à deux fenêtres fonctionne. S'il existe une constante c_3, une pénalité pen_h, une constante C et une contante $a = a(C) > 1$ telles que $\mathrm{pen}_{h^p(\beta,L)} = \rho_n(\beta)$ et

$$\left(\mathbb{P}_f\left\{\big|\mathring{f}^{(1)}(y) - \mathring{f}^{(2)}(y)\big| > C\,\sigma_p(h_1)\right\}\right)^{1/2} \le c_3 n^{-a}. \tag{1.5.4}$$

Alors d'après les cas traités ci-dessus, on a

$$\mathbb{E}_f\big|f^*(y) - f(y)\big| \le (2c_1 + c_2)\psi_n^p(\beta) + Kc_3n^{-a}, \quad \forall \beta, L \in [0, b[\times\mathbb{R}_+.$$

où $R_n = Kc_3n^{-a}$ est le terme de reste. Le proposition 1 est démontré.

Nous pouvons donner un schéma qui permet de contrôler la probabilité (1.5.4). Rappelons que nous sommes dans le cas où $h_1 < h_2 = h^p(\beta, L)$, et que $h_1 < h_2 \Rightarrow \sigma_p(h_1) > \sigma_p(h_2)$ si pen_h est décroissante en h. On peut écrire

$$\begin{aligned}
\mathbb{P}_f\left\{\big|\mathring{f}^{(1)}(y) - \mathring{f}_2(y)\big| > C\,\sigma_p(h_1)\right\} \;&\le\; \mathbb{P}_f\left\{\big|\mathring{f}^{(1)}(y) - f(y)\big| > \frac{C}{2}\,\sigma_p(h_1)\right\} \\
&\quad + \mathbb{P}_f\left\{\big|f(y) - \mathring{f}^{(2)}(y)\big| > \frac{C}{2}\,\sigma_p(h_1)\right\} \\
&= \sum_{j=1,2}\mathbb{P}_f\left\{\big|\mathring{f}^{(j)}(y) - f(y)\big| > \frac{C}{2}\,\sigma_p(h_j)\right\}.
\end{aligned}$$

Contrôler la probabilité (1.5.4) revient à contrôler les grandes déviations de l'estimateur $\mathring{f}^{(j)}$. Une grande partie de cette thèse est consacrée à la recherche de ce contrôle pour les estimateurs non-linéaires (voir Sections 2.1.3 et 2.2.2). Dans la suite, un exemple de contrôle des grandes déviations est donné pour l'estimateur de la moyenne locale dans le modèle de la régression gaussienne. La décomposition biais-variance joue un rôle particulièrement important, notamment sur le fait que le biais soit croissant et que la variance soit décroissante en h.

Grandes déviations : Moyenne locale. Considérons les observations $\mathcal{Z}_n$ dans le cadre de la régression gaussienne avec design fixe uniforme comme définie dans la section 1.2.4. Supposons que $nh^d \in \mathbb{N}^*$ pour simplifier les écritures et définissons l'estimateur de la moyenne locale :

$$\dot{f}^h = \frac{1}{nh^d} \sum_{X_i \in V_h(y)} Y_i, \quad \forall h \in [0,1].$$

Dans ce cas, l'estimation de la fonction f revient à approximer celle-ci par une constante sur un voisinage. Ici, nous prendrons $b = 1$, si $\beta \in]0,1]$, les fonctions $f \in \mathbb{H}(\beta, L)$ sont approchées par des polynômes de degré nul.

On est en présence d'un estimateur *linéaire*. Dans cette exemple la décomposition biais-variance est immédiate.

$$\left|\dot{f}^h(y) - f(y)\right| \le Ldh^\beta + \left|\frac{1}{nh^d} \sum_{X_i \in V_h(y)} \xi_i\right|, \quad \xi_i \sim \mathcal{N}(0, \sigma^2).$$

Supposons ici que la fenêtre $h \le h^p(\beta, L)$. Ainsi, par croissance du biais, décroissance de la variance et compromis biais-variance, on a

$$Ldh^\beta \le Ld\big(h^p(\beta,L)\big)^\beta = \sigma_p\big(h^p(\beta,L)\big) \le \sigma_p(h). \tag{1.5.5}$$

On obtient

$$\mathbb{P}_f\left\{\left|\dot{f}^h(y) - f(y)\right| > \frac{C}{2}\,\sigma_p(h)\right\} \le \mathbb{P}_f\left\{\left|\frac{1}{nh^d} \sum_{X_i \in V_h(y)} \xi_i\right| > \frac{C-2}{2}\,\sigma_p(h)\right\}.$$

Dans le cas gaussien $\gamma = 2$ ainsi il existe une constante $c_5 = c_5(\sigma^2) > 0$ telle que

$$\mathbb{P}_f\left\{\left|\frac{1}{nh^d} \sum_{X_i \in V_h(y)} \xi_i\right| > \frac{C-2}{2}\,\sigma_p(h)\right\} \le c_5 \exp\left\{-\left(\frac{C-2}{2}\,\mathrm{pen}_h\right)^2\right\}.$$

Cette inégalité s'obtient en intégrant la probabilité. Si nous choisissons la pénalité $\mathrm{pen}_h = \sqrt{1 + \ln \frac{h_{max}}{h}}$, on a

$$\mathbb{P}_f\left\{\left|\dot{f}^h(y) - f(y)\right| > \frac{C}{2}\,\sigma_p(h)\right\} \le c_5 \left(\frac{h_{max}}{h}\right)^{-\left(\frac{C-2}{2}\right)^2}.$$

Si l'on choisit h, h_{max} telles que $h_{max}h^{-1} \ge n^{c_6}$, on obtient le résultat avec $C = 2\sqrt{\frac{2}{c_6}} + 2$. ∎

Nous venons de voir que la méthode de Lepski repose sur des inégalités exponentielles de concentration, qui permettent de contrôler les grandes déviations. Ceci a été une difficulté importante qu'il a fallu surmonter dans cette thèse, notamment pour les estimateurs non-linéaires. Regardons maintenant la règle de sélection générale pour une grille finie d'estimateurs $\mathring{f}^{(k)}$.

Règle de Sélection Générale

Reconsidérons la famille des estimateurs $\mathcal{F}$. Nous pouvons maintenant donner la procédure de Lepski établie à partir de la règle de sélection à deux fenêtres. On prend $f^*(y) = \mathring{f}^{(\hat{k})}(y)$, où $\mathring{f}^{(\hat{k})}(y)$ est sélectionné dans $\mathcal{F}$ suivant la règle suivante :

$$\hat{k} = \inf\left\{k = \overline{0, \mathbf{k}_n} : \ \left|\mathring{f}^{(k)}(y) - \mathring{f}^{(l)}(y)\right| \le C\sigma_p(h_l), \ \ l = \overline{k+1, \mathbf{k}_n}\right\}, \tag{1.5.6}$$

où $\mathbf{k}_n$ est définie dans (1.5.2), $\overline{0, \mathbf{k}_n} = 0, ..., \mathbf{k}_n$ et la constante C est à choisir comme dans le cas précédent. Pour des exemples, on peut regarder les sections 2.1 et 2.2. $\mathbf{k}_n$ et $\mathcal{F}$ sont définis respectivement par (1.5.2) et (1.5.3). Notons que la fenêtre h_l est une fonction décroissante en l et la variance pénalisée $\sigma_p(h_l)$ est croissante en l.

Interprétation de la règle de Lepski

La règle de Lepski peut être expliquée par la description suivante. Considérons les fenêtres $0 < h_{\min} \le h_{\mathbf{k}_n} < ... < h_0 = h_{\max}$, on en déduit l'inclusion $\emptyset \ne V_{h_{\mathbf{k}_n}}(y) \subset ... \subset V_{h_1}(y)$. Dans ce cas, la variance de notre estimateur $\mathring{f}^{(k)}$ est croissante $Var\big(\mathring{f}^{(k)}\big) \le Var\big(\mathring{f}^{(k+1)}\big)$, $k = 1, ..., \mathbf{k}_n - 1$. En d'autres termes, le bias est en général décroissant quand le voisinage diminue $V_{h_k}(y)$. La méthode de Lepski peut être interprétée comme une procédure de tests multiples où l'hypothèse, que f est une polynôme sur $V_{h_k}(y)$, qui est testé contre le fait que l'écart entre les estimateurs est important. Notons cette hypothèse $H_0(k)$: $f_\theta|_{V_{h_k}(y)} = f|_{V_{h_k}(y)}$, où θ est la suite des coefficients des polynômes de Taylor.

Supposons $H_0(\mathbf{k}_n)$ est vraie, nous testons successivement si $H_0(k-1)$ is acceptable à condition que l'hypothèse $H_0(l)$ soit vérifiée pour tout $l \ge k$. Une fois le test de $H_0(k-1)$ rejeté, nous choisissons l'estimateur $\breve{f}^{(k)}$ correspondant à la dernière hypothèse. Le point principal est donc de bien définir les tests d'acceptation pour $H_0(k-1)$: la méthode de Lepski accepte $H_0(k-1)$ si l'événement $\left|\mathring{f}^{(k-1)}(y) - \mathring{f}^{(l)}(y)\right| \le CS_n(l)$ est vrai pour tout $l \ge k-1$ avec des valeurs de seuillage $CS_n(l)$ appropriées. Cette quantité $CS_n(l)$ peut être vu comme l'écart-type pénalisé de $\mathring{f}^l$ à une constante C-près.

Implémentation de la règle de Lepski

Soit $A_{j,l} = \left\{\left|\mathring{f}^{(j)}(y) - \mathring{f}^{(l)}(y)\right| \le C\sigma_p(h_l)\right\}$. Pour l'implémentation, on constate qu'il est nécessaire de calculer tous les estimateurs $\mathring{f}^{(l)}$ de la grille $\mathcal{F}$. Ceci donne la règle suivante :

on dit qu'un estimateur $\mathring{f}^{(j)}$ est *admissible* si l'événement $\bigcap_{j<l\leq \mathbf{k}_n} A_{j,l}$ est réalisé.

1. *On commence par vérifier si l'estimateur $\mathring{f}^{(0)}$ est admissible, ,*
2. *si il est admissible, on choisit $f^* = \mathring{f}^{(0)}$, sinon on passe au suivant $\mathring{f}^{(1)}$.*
3. *on s'arrête au premier instant où l'estimateur $\mathring{f}^{(j)}$ est admissible et tous les précédents $\mathring{f}^{(0)}, ..., \mathring{f}^{(j-1)}$ ne le sont pas. Ainsi, on prend $f^* = \mathring{f}^{(j)}$ où $j = \hat{k}$ est la plus petite des valeurs admissibles.*

Donc, dès le début de l'algorithme, pour vérifier que $\mathring{f}^{(0)}$ est admissible ou non, on construit tous les estimateurs de la grille $\mathcal{F}$.

Pour contourner ce problème, on peut utiliser l'implémentation de Katkovnik [1999] avec les *ICI* (en anglais : *Intersection of Confidence Intervals*), qui ne nécessite pas de calculer tous les estimateurs $\mathring{f}^{(l)}$ de la grille $\mathcal{F}$. La règle ICI est la suivante :

Implémentation des ICI

1. *Classons par odre croissant les fenêtres $h_{\mathbf{k}_n} < h_{\mathbf{k}_n - 1} < ... < h_1 < h_0$.*
2. *Considérons les intersections des intervalles de confiance $\mathcal{I}_j = \bigcap_{i=\mathbf{k}_n}^{j} \mathcal{D}_i$ pour $j \in \{\mathbf{k}_n, \mathbf{k}_n - 1, ..., 1, 0\}$, avec*

$$\mathcal{D}_i = \left[\mathring{f}^{(j)}(y) - C\varsigma_j, \mathring{f}^{(j)}(y) + C\varsigma_j\right],$$

   $\varsigma_j = \mathrm{Std}\left(\mathring{f}^{(j)}(y)\right)$ est l'écart-type de $\mathring{f}^{(j)}(y)$.
3. *Soit j^+ l'entier tel que pour tout $j > j^+$ on a $\mathcal{I}_j \neq \emptyset$, $\mathcal{I}_{j^+} \neq \emptyset$ et $\mathcal{I}_{j^+-1} = \emptyset$. Ainsi, l'estimateur adaptatif est $\mathring{f}^{(j^+)}(y)$.*

Une illustration des ICI est donnée par la figure 1.10. On constate que si j diminue, alors l'écart-type associé ς_j diminue. Ainsi, les intervalles de confiance diminuent quand j diminue. L'idée générale, cachée derrière les ICI, est que l'estimation du biais est plus petite, à constante près, que la variance tant que les intersections sont non-vides. La règle s'arrête quand l'estimation du biais et la variance sont proches.

La figure 1.11 permet de constater quelle fenêtre la méthode a choisie. La fenêtre $h_{\mathbf{k}-1}$ est trop petit, elle n'est pas optimale, nous pouvons encore l'agrandir. On passe à la fenêtre supérieure, on voit que celle-ci convient. La fenêtre $h_{\mathbf{k}-3}$ est trop grande et implique que $\mathcal{I}_{\mathbf{k}-3}$ est vide.

Si $\mathring{f}^{(j)}$ est un estimateur à noyau avec un noyau K_h, l'écart-type ς_j peut facilement être calculé par $\varsigma\|K_{h_j}\|_2$ (norme ℓ_2 du noyau, ς est l'écart-type du bruit). Dans les autres cas, on peut remplacer ς_j par $\sigma_p(h_j)$.

Asymptotiquement, on peut démontrer que l'estimateur ICI-adaptatif est proche de f^*, $\mathring{f}^{(j^+)}(y) \approx f^*(y)$ (voir Goldenshluger et Nemirovski [1997]). Pour les calculs théoriques, nous utiliserons l'estimateur f^* choisi par la règle de Lepski. En pratique, nous utiliserons $\mathring{f}^{(j^+)}$ l'estimateur choisi par la règle des ICI.

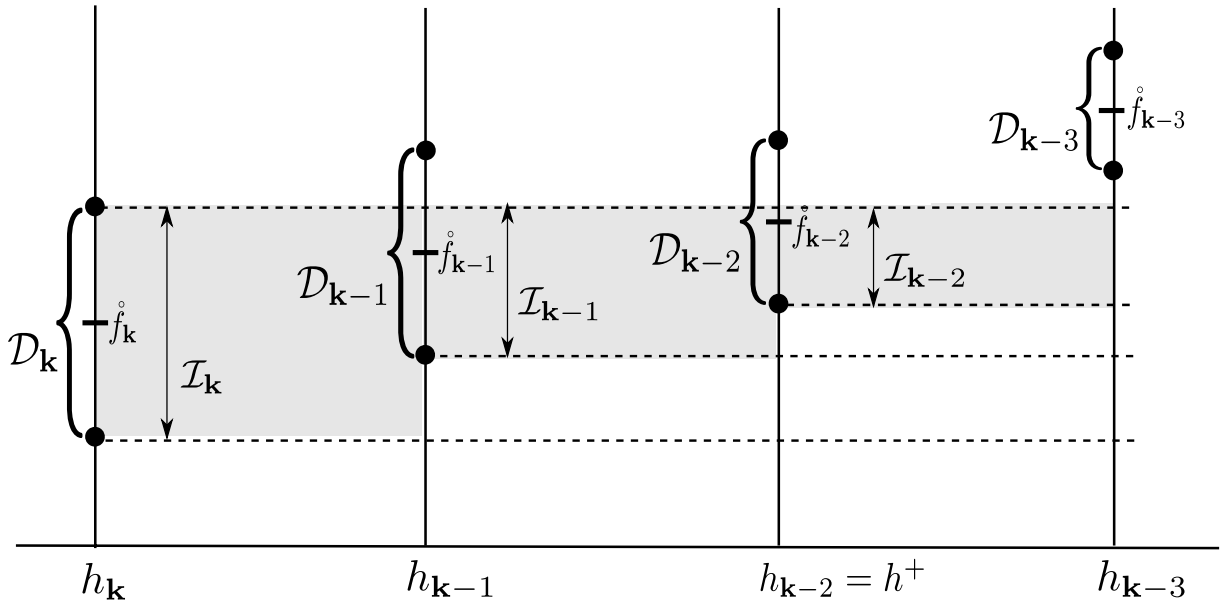


FIGURE 1.10 – Illustration de la règle ICI développée par Katkovnik [1999] ($\mathring{f}_j = \mathring{f}^{(j)}(y)$ et $h^+ = h_{j^+}$).

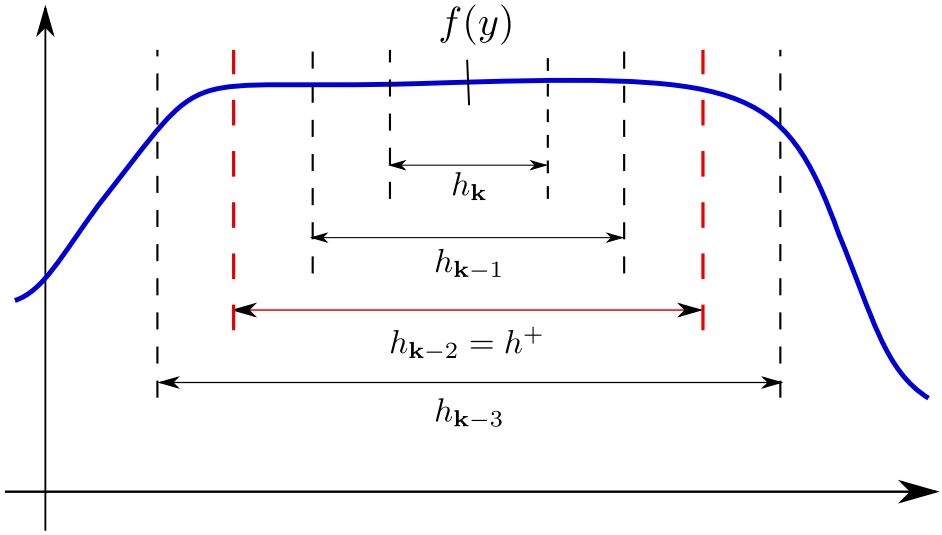


FIGURE 1.11 – Choix de la fenêtre $h^+ = h_{j^+}$ en traçant la fonction cible (exemple où la fonction est localement constante).

Comme nous venons de le voir, l'interprétation de la méthode de Lepski par les tests statistiques et son implémentation très simple par la méthode ICI ont fait le succès de cette méthode.

Nous pouvons démontrer que l'estimateur f^* est adaptatif.

Proposition 2. *Pour tout $\beta, L \in [0, b[\times\mathbb{R}_+$ on a*

$$\sup_{f\in\mathbb{H}_d(\beta,L)} \mathbb{E}_f\big|f^*(y) - f(y)\big| \le (Cc_1^* + c_2^*)\psi_n^p(\beta) + T_n,$$

où c_1^*, c_2^* sont des constantes positives et T_n est un terme de reste que l'on définira dans la preuve.

Dans la procédure (1.5.6), on peut remplacer C par une constante C' telle que $C' > C$. Le résultat du Proposition 2 reste vrai en remplaçant C par C' dans la borne supérieure (voir Preuve).

Idée de la preuve. Soit l'entier κ défini de la façon suivante.

$$2^{-\kappa}h_{\max} \le h^p(\beta, L) < 2^{-\kappa+1}h_{\max}.$$

On a aussi

$$\begin{aligned}\mathbb{E}_f\big|\mathring{f}^{(\hat{k})}(y) - f(y)\big| &\le \mathbb{E}_f\big|\mathring{f}^{(\hat{k})}(y) - f(y)\big|\mathbb{I}_{\hat{k}\le\kappa} + \mathbb{E}_f\big|\mathring{f}^{(\hat{k})}(y) - f(y)\big|\mathbb{I}_{\hat{k}>\kappa} \\ &:= R_1(f) + R_2(f).\end{aligned} \tag{1.5.7}$$

Premièrement, nous contrôlons R_1. Ainsi

$$\big|\mathring{f}^{(\hat{k})}(y) - f(y)\big| \le \big|\mathring{f}^{(\hat{k})}(y) - \mathring{f}^{(\kappa)}(y)\big| + \big|\mathring{f}^{(\kappa)}(y) - f(y)\big|.$$

La définition de $\hat{k}$ entraîne que

$$\big|\mathring{f}^{(\hat{k})}(y) - \mathring{f}^{(\kappa)}(y)\big|\mathbb{I}_{\hat{k}\le\kappa} \le C\sigma_p\big(h_\kappa\big).$$

Par définition de κ, l'estimateur $\mathring{f}^{(\kappa)}$ est le plus proche de l'estimateur minimax pénalisé dans la grille $\mathcal{F}$. Il est facile de démontrer qu'il existe une constante c_1^* telle que $\sigma_p\big(h_\kappa\big) \le c_1^*\psi^p(\beta)$ et une constante c_2^* telle que

$$\mathbb{E}_f\big|\mathring{f}^{(\kappa)}(y) - f(y)\big| \le c_2^*\psi^p(\beta).$$

D'où $R_1(f) \le (Cc_1^* + c_2^*)\psi^p(\beta)$.

Maintenant, il nous faut borner R_2. Appliquons l'inégalité de Cauchy-Schwarz

$$\begin{aligned}R_2(f) &= \sum_{k>\kappa}^{\mathbf{k}_n} \mathbb{E}_f\big|\mathring{f}^{(k)}(y) - f(y)\big| I_{\left[\hat{k}=k,G\right]} \\ &\le \sum_{k>\kappa} \big(\mathbb{E}_f\big|\mathring{f}^{(k)}(y) - f(y)\big|^2\big)^{1/2}\sqrt{\mathbb{P}_f\big\{\hat{k}=k\big\}}.\end{aligned}$$

Remarquons que l'indice $\hat{k}$ est aléatoire, nous le contrôlons à l'aide de la somme qui rend l'indexation déterminisme. Cela va nous permettre d'utiliser un contrôle des grandes déviations pour l'estimateur en question. Car si le paramètre l de l'estimateur localement polynômial $\mathring{f}^{(k)}$ est aléatoire (choisi à partir des données), alors cet estimateur n'est plus localement paramétrique. Ainsi le contrôle des grandes déviations est plus difficile à obtenir. Par exemple, les inégalités de concentration dit de *Bernstein* ou de *Berstein* (que l'on peut trouver dans Boucheron, Bousquet, et Lugosi [2004]) peuvent être utilisées pour les estimateurs linéaires seulement si la fenêtre est déterministe. Notons que pour tout $k \geq \kappa + 1$ et par définition de $\hat{k}$ (1.5.6)

$$\{\hat{k} = k\} = \cup_{l \geq k} \left\{ \left| \mathring{f}^{(k-1)}(y) - \mathring{f}^{(l)}(y) \right| > C\sigma_p(h_l) \right\}.$$

Notons que $\sigma_p(h_l)$ est strictement croissant en l, donc,

$$\begin{aligned} \{\hat{k} = k\} \quad \subseteq \quad & \left\{ \left| \mathring{f}^{(k-1)}(y) - f(y) \right| > 2^{-1} C\sigma_p(h_{k-1}) \right\} \\ & \cup \left[\cup_{l \geq k} \left\{ \left| \mathring{f}^{(l)}(y) - f(y) \right| > 2^{-1} C\sigma_p(h_l) \right\} \right]. \end{aligned}$$

On arrive à l'inégalité suivante : pour tout $k \geq \kappa + 1$

$$\begin{aligned} \mathbb{P}_f(\hat{k} = k) \quad \leq \quad & \mathbb{P}_f \left\{ \left| \mathring{f}^{(k-1)}(y) - \mathring{f}(y) \right| > 2^{-1} C\sigma_p(h_{k-1}) \right\} \\ & + \sum_{l \geq k} \mathbb{P}_f \left\{ \left| \hat{f}^{(l)}(y) - f(y) \right| > 2^{-1} C\sigma_p(h_l) \right\}. \end{aligned}$$

A ce moment de la preuve, nous pouvons donner une hypothèse pour finir la démonstration.

Hypothèses 2. *Pour tout $k \geq \kappa$, il existe une constante c_3^* telle que*

$$\mathbb{P}_f \left\{ \left| \hat{f}^{(k)}(y) - f(y) \right| > 2^{-1} C\sigma_p(h_k) \right\} \leq c_3^* \exp\left\{ - \left(\frac{C - c_4^*}{2} \operatorname{pen}_{h_k} \right)^{\gamma} \right\}.$$

Dans cette thèse, nous prendrons la pénalité $\operatorname{pen}_{h_k} = \left(1 + \ln\left[h_{\max} h_k^{-1}\right]\right)^{1/\gamma} = \left(1 + k \ln 2\right)^{1/\gamma}$. Avec l'hypothèse précédente, nous contrôlons

$$\mathbb{P}_f(\hat{k} = k) \leq c_5^* \exp\left\{ - \left(\frac{C - c_4^*}{2} \right)^{\gamma} (k-1) \ln 2 \right\},$$

où $c_5^* = c_3^* \sum_{l=0}^{+\infty} \exp\left\{ - \left(\frac{C - c_4^*}{2} \right)^{\gamma} l \ln 2 \right\}$. Ainsi, il est facile de contrôler le risque suivant. Il existe une constante c_6^* telle que

$$\left(\mathbb{E}_f \left| \mathring{f}^{(k)}(y) - f(y) \right|^2 \right)^{1/2} \leq c_6^* \sigma_p(h_k).$$

On obtient ce résultat en intégrant la probabilité de l'hypothèse 2. D'où

$$R_2(f) \leq \sum_{k>\kappa} c_6^* \sigma_p(h_k) \sqrt{c_5^*} \exp\left\{-\frac{1}{2}\left(\frac{C-c_4^*}{2}\right)^\gamma (k-1)\ln 2\right\}.$$

Avec le choix de $C = 2(2d)^{1/\gamma} + c_4^*$, on obtient

$$R_2(f) \leq T_n = \frac{1}{\left(nh_{\max}^d\right)^{1/\gamma}} \sum_{k>0} c_6^* \sqrt{c_5^*}(1+k\ln 2)^{1/\gamma} \exp\left\{-\frac{1}{2}\left(\frac{C-c_4^*}{2}\right)^\gamma (k-1)\ln 2\right\}.$$

Avec cette inégalité, l'assertion du proposition est démontrée.

La recherche, du contrôle des grandes déviations, est une partie technique mais indispensable pour démontrer les résultats d'adaptation. Cette thèse consacre une partie importante à la recherche de ce type d'inégalités notamment pour les estimateurs non-linéaires. De nombreux papiers sont consacrés à ces résultats dans le but de l'adaptation (voir notamment Talagrand [1994], Ledoux [1997], Massart [2000, 2007], Boucheron, Bousquet, et Lugosi [2004] et Goldenshluger et Lepski [2009b]).

Nous insistons sur le fait que le biais et la variance de notre estimateur doivent être respectivement croissant et décroissante (voir l'équation (1.5.5)). Pour les fonctions anisotropes, le biais n'est pas monotone. Dans ce sens, les articles de Kerkyacharian, Lepski, et Picard [2001], Klutchnikoff [2005], Goldenshluger et Lepski [2008] et Goldenshluger et Lepski [2009a] utilisent une méthode s'appuyant sur une comparaison des biais. La notion d'emboîtement est inévitable. En effet il est nécessaire de trouver un ordre sur l'ensemble des fenêtres pour que la variance soit monotone. En général, pour l'approche localement paramétrique, le nombre d'observations appartenant à $V_h(y)$ augmente quand la fenêtre augmente, d'où la diminution de la variance.

Dans cette procédure, nous avions un "point d'appui", $\mathring{f}^{(\kappa)}$ l'estimateur minimax pénalisé, qui est notre *oracle*. Ainsi, la méthode de Lepski peut donner des inégalités oracle (voir Goldenshluger et Lepski [2009a]), mais cela demande des résultats un peu plus techniques (bornes uniformes de fonctions aléatoires, voir Goldenshluger et Lepski [2009b]).

La méthode de Lepski (Lepski, Mammen, et Spokoiny [1997]) nous a permis de construire des estimateurs *bayésien et de Huber adaptatifs optimaux*. Nous présentons dans le chapitre suivant les résultats obtenus dans cette thèse notamment grâce à cette méthode. Dans la section 2.3, la performance de cette méthode est évaluée (par rapport à l'oracle). Une illustration de la calibration de la constante C est aussi présentée.

Chapitre 2

Résultats et Perspectives

Après avoir défini les modèles, les nouvelles méthodes d'estimation et la procédure adaptative, nous présentons les différents théorèmes qui mesurent les performances des estimateurs proposés. D'un point vue minimax dans un premier temps, nous développons ensuite des procédures adaptatives basée sur l'idée de Lepski. Comme nous l'avons déjà souligné ci-dessus, nos résultats sont déduits du contrôle des grandes déviations de nos différents estimateurs. Ces inégalités font l'objet de sous-sections pour chaque méthode d'estimation, et peuvent être étudiées de façon indépendante des autres résultats. En fin de paragraphe, nous parlerons des suites à donner à ces travaux, ainsi que des problèmes ouverts.

2.1 Approche Bayésienne

Nous donnons les résultats obtenus pour l'estimateur bayésien dans le modèle de la régression générale. Tout d'abord nous précisons quelques hypothèses : la vitesse de convergence minimax est supposée *connue* $\varphi_{n,\gamma}(\beta) = n^{-\frac{\beta}{\gamma\beta+d}}$ sur les espaces de Hölder $\mathbb{H}_d(\beta, L)$, où $1 \leq \gamma \leq 2$ est une constante qui dépend de la densité $g(.,.)$ des observations. Une propriété remarquable de l'estimateur bayésien est qu'il s'adapte à la densité du modèle pour être optimal au sens des vitesses de convergence minimax. Comme tout estimateur localement paramétrique, le choix de la fenêtre est primordial pour construire un estimateur qui ne dépend pas de la fonction à estimer. Nous proposons une procédure adaptative pour le choix de la fenêtre, basée sur l'idée de Lepski, qui construit un estimateur adaptatif optimal. Nous verrons quelques exemples d'optimalité pour certains modèles (section 4.3). Comme nous l'avons vu dans la section 1.5.2, le contrôle des grandes déviations pour l'estimateur bayésien indexé par h doit être obtenu pour tout h. Ici, l'estimateur bayésien n'est pas linéaire et on ne connaît pas sa forme explicite. Nous utiliserons le fait qu'il minimise un critère particulier pour contrôler les grandes déviations. Une fois les résultats énoncés pour le modèle de régression générale, nous étudions quelques exemples comme la régression gaussienne et la régression inhomogène de Poisson avec la vitesse $\varphi_{n,2}(\beta)$, la régression α avec $\varphi_{n,1+2\alpha}(\beta)$ et

la régression multiplicative uniforme avec $\varphi_{n,1}(\beta)$.

Nous proposons une procédure adaptative pour l'estimateur bayésien. Cette procédure permet un choix à partir des observations de la fenêtre d'estimation (en anglais : bandwidth). Les résultats obtenus sont de type minimax adaptatifs pour le risque ponctuel. En particulier notre estimateur adaptatif atteint la vitesse $\varphi_{n,\gamma}(\beta)$, $\forall \beta \in]0, b]$, nous verrons plusieurs modèles dans lesquels cette vitesse est optimale. Notons toutefois que l'adaptation pour l'approche bayésienne est déjà apparue, par exemple dans les travaux de Ghosal, Lember, et Van der Vaart [2008] et Van der Vaart et Van Zanten [2009].

2.1.1 Recherche de la Vitesse Minimax

Rappelons le cadre dans lequel on travaille.

1. Le modèle est celui de la régression générale défini dans (1.2.2) avec une densité $g(.,.)$,
2. $f \in \mathbb{H}_d(\beta, L, M)$ (Définition 1),
3. on utilise l'estimateur bayésien $\hat{f}$ défini dans (1.3.7), avec la constante m à choisir qui correspond à l'inverse de la puissance du rapport de vraisemblance,
4. on suppose que la vitesse minimax est connue et égale à $\varphi_{n,\gamma}(\beta) = n^{-\frac{\beta}{\gamma\beta+d}}$ avec γ connue.

Pour obtenir la borne supérieure du risque maximal (Définition 6), nous avons besoin des hypothèses suivantes. Pour cela , on introduit les notations suivantes. Pour tout $\theta \in \Theta(M)$, soit $U_n = N_h \big[\Theta\big(M\big) - \theta\big]$. Donnons aussi $N_h = \big(nh^d\big)^{1/\gamma}$ où $\gamma \geq 1$ et $\forall u \in U_n$ le pseudo rapport de vraisemblance

$$Z_{n,\theta}(u) = \frac{L_h\big(\theta + u\, N_h^{-1}, \mathcal{Z}_n\big)}{L_h\big(\theta, \mathcal{Z}_n\big)}. \tag{2.1.1}$$

Soit $\mathcal{H}_n, n > 1$ le sous-intervalle suivant de $(0, 1)$.

$$\mathcal{H}_n = [h_{\min},\, h_{\max}]\,, \quad h_{\min} = \big(\ln n\big)^{\frac{1}{\gamma d+d^2}} n^{-1/d}, \quad h_{\max} = (\ln n)^{-\frac{1}{\gamma b+d}}\,. \tag{2.1.2}$$

Dans la suite, nous considérons seulement les valeurs de h appartenant à $\mathcal{H}_n$. Soit $f_\theta(x)$, avec $\theta \in \Theta\big(M\big)$, le polynôme local d'approximation de f sur $V_h(y)$ et soit b_h l'erreur d'approximation correspondante, i.e.

$$b_h = \sup_{x \in V_h(y)} \big|f_\theta(x) - f(x)\big|. \tag{2.1.3}$$

Finalement, définissons $\mathcal{N}(h) = \big(b_h \times N_h\big)^\gamma$ et

$$\mathcal{E}_h = \exp\left\{\mathcal{N}(h)\right\}. \tag{2.1.4}$$

Soit le sous-intervalle $\Gamma_\delta \subseteq \Theta\big(M\big)$ tel que $\int_{\Gamma_\delta} du = \delta^{D_b}$, $\delta > 0$.

Hypothèses 3. *On suppose qu'il existe plusieurs constantes $\tau, m > 0$ et c_1, c_2, c_3, C_1, C_2, s_1, $s_2 > 0$ telles que pour tout $n > 1$, $f \in \mathbb{H}_d(\beta, L, M)$, $\theta \in \Theta\big(M\big)$ et $h \in \mathcal{H}_n$, on a*

1. $\displaystyle\int_{\Gamma_\delta} \mathbb{E}_f \left[1 - Z_{n,\theta}^{1/m}(u)\right]_+ du \le C_1\, \mathcal{E}_h^{c_1} \delta^{D_b+\tau}, \quad \forall \delta < s_1,$

2. $\mathbb{E}_f Z_{n,\theta}^{1/m}(u) \le C_2\, \mathcal{E}_h^{c_2}\ \exp\left\{-c_3\, \|u\|_1^\gamma\right\}, \ \ \forall u \in U_n : \ \|u\|_1 \ge s_2\, \mathcal{N}(h).$

où $a_+ = \max(a, 0)$, $a \in \mathbb{R}$.

Remarque 3. *On suppose aussi connues les constantes c_3, s_2, τ, m and γ, l'existence des autres constantes est suffisante. Notons que le processus $Z_{n,\theta}(.)$ n'est pas toujours défini, par exemple si la mesure de probabilité $\mathbb{P}_{f_{\theta+uN_h^{-1}}}$ n'est pas absolument continue par rapport à $\mathbb{P}_{f_\theta}$. Dans ce cas, le problème est ouvert.*

Par le passé, Has'minskii et Ibragimov [1981] ont énoncé des hypothèses similaires pour les modèles paramétriques. Ils utilisent l'estimateur bayésien $\int ||t-u||L(u)du$ mais sans la puissance $1/m$. Ainsi, les hypothèses ci-dessus sont plus faibles que celles de Has'minskii et Ibragimov [1981] et Chichignoud [2010a] (voir Chapitre 4). Ces hypothèses sont données par Chichignoud [2010b] (Chapitre 3). La continuité du processus $Z_{n,\theta}(.)$, et donc de la vraisemblance, n'est pas nécessaire pour l'approche bayésienne. En revanche, $Z_{n,\theta}$ est défini comme le rapport de deux mesures de probabilités, il est donc nécessaire que $\mathbb{P}_{f_{\theta+uN_h^{-1}}}$ soit absolument continue par rapport à $\mathbb{P}_{f_\theta}$. Le prochain théorème montre qu'avec un choix judicieux de la fenêtre h, l'estimateur bayésien local est optimal d'un point de vue minimax sur les espaces de Hölder. Posons $\bar{h} = \big(L^\gamma n\big)^{-\frac{1}{\gamma\beta+d}}$ et soit $\bar{f}^{\bar{h}}(y) = \hat{\theta}_{0,\dots,0}\big(\bar{h}\big)$ donné par (1.3.4), (1.3.6) et (1.3.7) avec $h = \bar{h}$.

Théorème 1. *Soient $\beta > 0$, $L > 0$ et $M > 0$ fixées. Supposons que les hypothèses 3 sont vérifiées. Alors, il existe une constante C_* telle que pour tout $n \in \mathbb{N}^*$ satisfaisant $N_{\bar{h}} \ge 1$,*

$$\varphi_{n,\gamma}^{-q}(\beta) R_{n,q}\Big[\bar{f}^{\bar{h}}(y), \mathbb{H}_d(\beta, L, M)\Big] \le C_*, \quad \forall q \ge 1.$$

Remarque 4. *Sous les hypothèses 3, nous savons construire un estimateur minimax. Nous verrons dans les sections 2.1.4 et 3.4 dans quels modèles ces hypothèses sont vérifiées. Les estimateurs de type bayésien ou de Pitman ont été introduits dans le livre de Has'minskii et Ibragimov [1981] pour l'estimation paramétrique, dans le but d'atteindre la vitesse optimale pour différents modèles où la vitesse varie. Dans la preuve (section 3.5), la constante C_* est exprimée de façon explicite. Notons que la borne inférieure n'est pas calculée car nous avons supposé la vitesse minimax connue. Dans les exemples (section 2.1.4), les bornes inférieures du risque minimax sont calculées.*

2.1.2 Procédure Adaptative

Cette section est vouée à l'estimation adaptative sur la collection de classes $\Big\{\mathbb{H}_d(\beta,L,M)\Big\}_{\beta,L,M}$. Nous n'allons pas imposer de restriction sur les valeurs possibles de L, M. En effet, L, M joue un rôle secondaire en intervenant dans les constantes, à l'inverse de β qui a une influence sur la vitesse. Nous supposerons que $\beta \in (0, b]$, où b, est un entier choisi arbitrairement *a priori*. Notons aussi que la constante M peut être estimée par un estimateur non nécessairement optimal (voir Chichignoud [2010a] ou Chapitre 4, Section 4.3). En revanche, il n'existe aucune méthode capable d'estimer β et L.

Soit Φ la famille de normalisations suivante :

$$\phi_{n,\gamma}(\beta) = \left(\frac{\rho_{n,\gamma}(\beta)}{n}\right)^{\frac{\beta}{\gamma\beta+d}}, \quad \rho_{n,\gamma}(\beta) = \left[1 + \frac{\gamma(b-\beta)}{(\gamma b+d)(\gamma\beta+d)}\ln n\right]^{\frac{1}{\gamma}}, \quad \beta \in (0,b].$$

On remarque que $\phi_{n,\gamma}(b) = \varphi_{n,\gamma}(b)$ et $\rho_{n,\gamma}(\beta) \sim \big(\ln n\big)^{\frac{1}{\gamma}}$ pour tout $\beta \neq b$. On constate une perte dans la vitesse adaptative. Ceci est en général toujours vrai pour l'estimation ponctuelle (voir notamment Lepski [1990], Tsybakov [1998], Klutchnikoff [2005] et Chichignoud [2010a]). Nous verrons que cette perte est nécessaire pour contrôler les grandes déviations pour l'*estimateur bayésien local*.

Construction de l'estimateur Φ-adaptatif. Comme mentionné dans la section 1.5.2, la construction de notre procédure d'estimation comprend deux étapes. Premièrement, nous définissons la famille des estimateurs bayésiens locaux. Ensuite, s'appuyant sur la méthode de Lepski, nous proposons une sélection adaptative pour cette famille.

1ère étape : Collection d'estimateurs bayésiens locaux. Posons

$$h_k = 2^{-k}h_{\max}, \quad k = 0,\dots,\mathbf{k}_n,$$

où $\mathbf{k}_n$ est le plus grand entier tel que $h_{\mathbf{k}_n} \in \mathcal{H}_n$ défini dans (2.1.2). Soit

$$\hat{\mathcal{F}} = \Big\{\hat f^{(k)}(y) = \hat\theta_{0,\dots,0}(h_k), \quad k = 0,\dots,\mathbf{k}_n\Big\},$$

où $\hat\theta_{0,\dots,0}(h_k)$ est donné par (1.3.4), (1.3.6) et (1.3.7) avec $h = h_k$.

2ème étape : Sélection parmi la collection $\hat{\mathcal{F}}$. On pose $\hat f^*(y) = \hat f^{(\hat k)}(y)$, où $\hat f^{(\hat k)}(y)$ est sélectionné dans $\hat{\mathcal{F}}$ avec la règle de décision suivante :

$$\hat k = \inf\Big\{k = \overline{0,\mathbf{k}_n} : \ \big|\hat f^{(k)}(y) - \hat f^{(l)}(y)\big| \le CS_n\big(l\big), \ \ l = \overline{k+1,\mathbf{k}_n}\Big\}. \tag{2.1.5}$$

Ici, on a utilisé les notations suivantes, pour $q \ge 1$:

$$C = \left(s_2^\gamma \vee \frac{2^{2\gamma+2}(\gamma+dq)}{c_3\gamma\left(1\wedge\tau\, D_b^{-1}\right)}\right)^{\frac{1}{\gamma}}, \quad S_n(l) = \left[\frac{1+l\ln 2}{n\big(h_l\big)^d}\right]^{\frac{1}{\gamma}}, \quad l = 0,1,\dots,k_n. \tag{2.1.6}$$

Théorème 2. *Soit $b > 0$ fixé. Supposons que les hypothèses 3 soient vérifiées, alors pour tout $\beta \in (0, b]$, $L > 0$, $M > 0$ et $q \geq 1$*

$$\limsup_{n\to\infty} \phi_{n,\gamma}^{-q}(\beta) R_{n,q}\Big[\hat{f}^*(y), \mathbb{H}_d(\beta, L, M)\Big] < \infty.$$

Remarque 5. *Ce théorème correspond au théorème 5 dans le chapitre 3. L'assertion de ce théorème montre que l'estimateur proposé $\hat{f}^*(y)$ est Φ-adaptatif. Cela implique que la famille de normalisations Φ est admissible. Par exemple, nous pouvons démontrer l'optimalité de Φ dans la régression multiplicative uniforme avec le critère de Klutchnikoff [2005] (voir Section 4.6). Nous avons choisi d'écrire le théorème de façon asymptotique pour simplifier la présentation, au vue de la preuve (section 3.5) nous pourrions écrire la borne supérieure sous la forme $\phi_{n,\gamma} + R_n$ comme présentée dans la section 1.5.2. La connaissance des constantes s_2, γ, c_3 est nécessaire pour construire la constante C intervenant dans la procédure. En pratique, on peut calibrer la constante C sur des jeux de données et ainsi éviter ce problème.*

2.1.3 Grandes Déviations

Nous avions précisé dans la section 1.5.2 que le contrôle des grandes déviations est indispensable pour obtenir l'adaptation. Ainsi, nous avons pu établir un nouveau type d'inégalités exponentielles. Il existe ε_0, ω_1, ω_2 et ω_3 des constantes positives. Pour tout $n > 1$, $h \in \mathcal{H}_n$ et $f \in \mathbb{H}_d(\beta, L, M)$, alors $\forall \varepsilon \geq \varepsilon_0 \times \mathcal{N}(h)$

$$\mathbb{P}_f\left(N_h \big|\hat{f}^h(y) - f(y)\big| \geq \varepsilon\right) \leq \omega_1 \mathcal{E}_h^{\omega_2} \exp\left\{-\omega_3 \varepsilon^\gamma\right\},$$

où l'estimateur bayésien $\hat{f}^h(y)$ est donné par (1.3.4), (1.3.6) et (1.3.7). Ce résultat est énoncé dans la proposition 3, la preuve est disponible dans la section 3.5.4. La nouveauté ici est la puissance γ dans l'exponentielle. En effet la puissance conditionne la vitesse $\phi_{n,\gamma}(\beta)$ (voir aussi Has'minskii et Ibragimov [1981]). A noter aussi l'apparition du terme $\mathcal{E}_h = \exp\{b_h N_h\}$. Ce terme peut éventuellement tendre vers l'infini si le nombre d'observations augmente. Pour le contrôler, nous utiliserons le compromis biais-variance, i.e. $\mathcal{N}(h) = b_h N_h \leq L d h^\beta \times \left(nh^d\right)^{1/\gamma}$, nous pouvons voir ce terme comme le rapport biais/variance. Pour l'estimateur minimax, la borne supérieure du risque s'obtient par intégration de cette probabilité.

Dans la suite, nous donnons quelques idées essentielles de la preuve. Notamment comment utiliser les hypothèses 3 et le critère bayésien.

Idée de la preuve. La définition de $\hat{\theta}(h)$ et $\theta = \theta(f, y, h)$ les coefficients du polynôme de Taylor associé à f, impliquent que $\forall \varepsilon > 0$

$$\begin{aligned} \mathbb{P}_f\left(N_h \big|\hat{f}^h(y) - f(y)\big| \geq \varepsilon\right) &\leq \mathbb{P}_f\left(N_h \big|\hat{\theta}_0(h) - \theta_0\big| \geq \varepsilon\right) \\ &\leq \mathbb{P}_f\left(N_h \big\|\hat{\theta}(h) - \theta\big\|_1 \geq \varepsilon\right). \end{aligned} \tag{2.1.7}$$

Quelques remarques sont à faire. Premièrement, par définition on a $\theta \in \Theta(M)$. Rappelons que $\hat{\theta}(h)$ minimise π_h défini par (1.3.6) et donc, l'inclusion suivante est vraie si $\hat{\theta}(h) \in \Theta(M)$.

$$\left\{ N_h \left\| \hat{\theta}(h) - \theta \right\|_1 \geq \varepsilon \right\} \subseteq \left\{ \inf_{N_h \|t-\theta\|_1 \geq \varepsilon} \pi_h(t) \leq \pi_h(\theta) \right\}. \tag{2.1.8}$$

En outre,

$$\begin{aligned} \pi_h(t) &= N_h^{-1} \int_\Theta \|N_h(t-u)\|_1 \left[L_h\big(u, \mathcal{Z}_n\big) \right]^{1/m} du \\ &= N_h^{-D_b-1} \int_{U_n} \|N_h(t-\theta) - u\|_1 \left[L_h\big(\theta + uN_h^{-1}, \mathcal{Z}_n\big) \right]^{1/m} du \\ &= N_h^{-D_b-1} \left[L_h\big(\theta, \mathcal{Z}_n\big) \right]^{1/m} \int_{U_n} \|N_h(t-\theta) - u\|_1 \, Z_{n,\theta}^{1/m}(u) du. \end{aligned}$$

D'où, $\tau_n = N_h\big(\hat{\theta}(h) - \theta\big)$ est le minimiseur de

$$\chi_n(s) = \int_{U_n} \left\| s - u \right\|_1 \frac{Z_{n,\theta}^{1/m}(u)}{\int_{U_n} Z_{n,\theta}^{1/m}(v) dv} du,$$

et on obtient avec (2.1.7) et (2.1.8) pour tout $\varepsilon > 0$

$$\mathbb{P}_f \left\{ \left\| N_h(\hat{\theta}(h) - \theta) \right\|_1 > \varepsilon \right\} \leq \mathbb{P}_f \left\{ \inf_{\|s\|_1 > \varepsilon} \chi_n(s) \leq \chi_n(0) \right\}. \tag{2.1.9}$$

Soient $\nu > 0$ et $0 < r < \varepsilon/3$ deux paramètres à choisir, après quelques minorations et majorations de $\chi_n(.)$, on obtient

$$\begin{aligned} \mathbb{P}_f \left\{ \chi_n(0) - \inf_{\|s\|_1 > \varepsilon} \chi_n(s) > 0 \right\} \leq\ & 2\mathbb{P}_f \left\{ \int_{U_n \cap (\|u\|_1 > r)} \|u\|_1 Z_{n,\theta}^{1/m}(u) du > r\,\nu/4 \right\} \\ & + 2\mathbb{P}_f \left\{ \int_{U_n} Z_{n,\theta}^{1/m}(v) dv < \nu/2 \right\}. \end{aligned}$$

En utilisant l'inégalité de Markov, le premier terme peut être contrôlé par une inégalité exponentielle avec l'hypothèse 3.2. L'hypothèse 3.1 est utilisée pour contrôler le deuxième terme en retranchant $v Z_{n,\theta}^{1/m}(0)$ dans l'inégalité. Ainsi, le contrôle des grandes déviations avec la puissance γ est obtenu pour l'estimateur *bayésien* dans le modèle de *régression générale*. Plus de détails sont donnés dans la preuve 3.5.4 de la proposition 3.

2.1.4 Exemples de Modèles avec des Vitesses Différentes

Dans cette partie, nous précisons les modèles de régression (gaussienne, inhomogène de Poisson, α et multiplicative uniforme) où les hypothèses 3 sont vérifiées. Pour chaque modèle,

on construit un *estimateur bayésien local minimax*. A l'aide de la procédure adaptative, nous donnons un *estimateur bayésien adaptatif optimal* pour chaque modèle.

Nous nous intéressons à trouver des modèles dans lequels les estimateurs linéaires ne sont pas optimaux (au sens minimax et adaptation minimax). Par exemple, dans les régressions α et multiplicative uniforme, les vitesses minimax sont respectivement $\varphi_{n,1+2\alpha}(\beta)$ et $\varphi_{n,1}(\beta)$ et donc bien meilleures que $\varphi_{n,2}(\beta)$ la vitesse des estimateurs linéaires. Nous pouvons dire que la famille des estimateurs bayésiens indexés par la fenêtre est plus "riche" que celle des estimateurs linéaires.

Nous ne précisons pas encore, à ce stade, comment vérifier les hypothèses 3. Ce travail est présenté dans la section 3.4. Les démonstrations doivent être faites à la "main", i.e. pour une densité $g(.,.)$ donnée, on doit rechercher les constantes qui vérifient les hypothèses 3.

Régression Gaussienne. La régression gaussienne est très connue en statistiques non-paramétriques. Beaucoup d'estimateurs linéaires ont déjà été développés et sont optimaux. On veut vérifier que notre estimateur bayésien fait aussi bien que les estimateurs linéaires dans ce cadre très classique. Rappelons le modèle additif gaussien

$$Y_i = f(X_i) + \xi_i, \quad \xi_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, ..., n,$$

où $f \in \mathbb{H}(\beta, L, M)$ et $\sigma > 0$. Ici la densité des observations s'écrit

$$g(., f(X_i)) = g_\xi(. - f(X_i)) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{\left(. - f(X_i)\right)^2}{2\sigma^2}\right\}.$$

On peut vérifier les hypothèses 3 avec $\gamma = 2$. Ainsi, l'estimateur bayésien local atteint les vitesses minimax $\varphi_{n,2}(\beta) = n^{-\frac{\beta}{2\beta+d}}$ et adaptative $\phi_{n,2}(\beta)$ pénalisée (voir Thérorème 6) avec un $\sqrt{\ln n}$ qui est optimale pour la régression gaussienne. Le choix des fenêtres minimax et adaptatives est donné respectivement dans les sections 2.1.1 et 2.1.2.

Régression Inhomogène de Poisson. Nous sommes ici dans le cas discret, avec une densité

$$g(k, f(X_i)) = \mathbb{P}_f(Y_i = k) = \frac{[f(X_i)]^k}{k!} \exp\left\{-f(X_i)\right\}, \quad k \in \mathbb{N}, \quad f \in \mathbb{H}_d(\beta, L, M, A).$$

Dans ce cadre, on peut dire que $Y_i \sim \mathcal{P}\big(f(X_i)\big)$. Les hypothèses 3 sont vérifiés avec $\gamma = 2$ (Lemma 2). Ainsi, l'estimateur bayésien local atteint les vitesses minimax $\varphi_{n,2}(\beta) = n^{-\frac{\beta}{2\beta+d}}$ et adaptative $\phi_{n,2}(\beta)$ pénalisée avec un $\sqrt{\ln n}$ (Thérorème 7). Les choix des fenêtres minimax et adaptatives sont données respectivement dans les sections 2.1.1 et 2.1.2. Les vitesses de convergence sont les mêmes que celles de la régression gaussienne. Ce qui est naturel par équivalence asymptotique entre ces deux modèles (voir Anscombe [1948]).

Régression α. La régression α justifie la construction de l'estimateur bayésien. En effet, la vitesse de convergence dépend de α et s'améliore dès que $\alpha < 1/2$. Rappelons le modèle

$$Y_i = f(X_i) + \epsilon_i, \quad i = 1, ..., n,$$

où ϵ_i a pour densité $g_\alpha(x) = C(\alpha) \exp\{-|x|^\alpha\}$ avec $0 < \alpha < 1/2$. $C(\alpha)$ est une constante choisie pour que g_α soit bien une densité. Ici le design X_i est déterministe uniformément réparti sur $[0,1]^d$ et $f \in \mathbb{H}_d(\beta, L, M)$.

Nous démontrons que l'estimateur bayésien atteint la vitesse $\varphi_{n,1+2\alpha}(\beta) = n^{-\frac{\beta}{(1+2\alpha)\beta+d}}$ si l'on connaît β. La procédure adaptative, donnée dans la section 2.1.2, permet de construire un estimateur bayésien adaptatif qui atteint la vitesse $\phi_{n,1+2\alpha}(\beta)$ pénalisée avec un $\big(\ln n\big)^{\frac{1}{(1+2\alpha)}}$. On remarque que pour $\alpha < 1/2$, on a la vitesse de convergence $\varphi_{n,1+2\alpha}(\beta) < \varphi_{n,2}(\beta)$. Ainsi, dans ce modèle, l'estimateur bayésien est "meilleur" (converge plus vite) que les estimateurs linéaires.

Has'minskii et Ibragimov [1981] démontrent, dans le cas paramétrique, que pour $\alpha > 1/2$ la vitesse est $n^{-1/2}$ et pour $\alpha = 1/2$ on a la vitesse $\big(n \ln n\big)^{-1/2}$. Donc $1/2$ est la frontière avec la "pire" vitesse $\varphi_{n,2}(\beta)$. On remarque que si $\alpha \approx 0$ alors $\varphi_{n,1+2\alpha}(\beta) \approx \varphi_{n,1}(\beta)$ où $\varphi_{n,1}(\beta)$ est apparemment la "meilleure" vitesse que l'on peut atteindre en estimation.

Régression Multiplicative Uniforme. Le modèle est

$$Y_i = f(X_i) \times U_i, \quad i = 1, \ldots, n,$$

où $f \in \mathbb{H}_d(\beta, L, M, A)$. Les variables aléatoires $(U_i)_{i\in 1,\ldots,n}$ sont supposées indépendantes et uniformément distribuées sur $[0,1]$ et les points du design $(X_i)_{i\in 1,\ldots,n}$ sont déterministes.

L'étude de ce modèle est l'objet du chapitre 4. L'approche bayésienne développée ne peut pas s'appliquer directement. Premièrement, nous avons construit deux estimateurs $\hat{A}$ et $\hat{M}$ de A et de M, et introduit l'estimateur bayésien construit sur l'ensemble des coefficients aléatoires. Deuxièmement, le processus $Z_{n,\theta}$ n'est pas toujours défini (les mesures de probabilité ne sont pas absolument continues). Il nous a fallu choisir de façon judicieuse le polynôme d'approximation f_θ avec la condition $f_\theta \geqq f$.

Ainsi nous prouvons que la vitesse minimax est $\varphi_{n,1}(\beta) = n^{-\frac{\beta}{\beta+d}}$ (Théorème 10) et que notre estimateur bayésien est minimax (voir Théorème 11, Section 4.2). Dans le cadre adaptatif, nous donnons une procédure plus spécifique que celle présentée ci-dessus, qui nous permet de trouver un estimateur $\phi_{n,1}(\beta)$-adaptatif (Théorème 14) où la pénalité est en $\ln n$. Nous démontrons à l'aide du critère de Klutchnikoff [2005] que cette famille de normalisations est adaptative optimale (voir Théorème 13). On rappelle qu'il y a toujours un prix à payer pour l'adaptation en estimation ponctuelle.

On rappelle que pour utiliser l'estimateur bayésien, les hypothèses 3 doivent être vérifiées, en particulier pour les exemples ci-dessus. On peut trouver plus de détails dans la section 3.4 (Chapitre 3).

2.2 Critère de Huber

Cette partie présente un estimateur "robuste" (peu sensibles aux valeurs extrêmes, en anglais *outliers*) de type M-estimateur que nous appelons estimateur de *Huber*. Comme nous l'avons vu dans la section précédente, l'estimateur bayésien recherche la vitesse optimale pour un modèle donné. L'inconvénient de l'approche bayésienne est qu'il faut supposer connue la densité des observations $g(.,.)$ pour vérifier les hypothèses 3. Nous proposons un estimateur qui ne dépend pas de la densité g_ξ des observations (avec g_ξ vérifiant les hypothèses 1). En revanche, l'estimateur de Huber n'est pas toujours optimal (par exemple régression α ou régression additive uniforme).

Rappelons dans quel cadre on travaille.

1. Le modèle est celui de la régression additive avec design aléatoire (indépendant du bruit) défini dans la section 1.2.3 avec une densité $g_\xi(.)$ vérifiant les hypothèses 1,

$$Y_i = f(X_i) + \xi_i, \quad i = 1, ..., n,$$

2. $f \in \mathbb{H}_d(\beta, L, M)$ (Définition 1),
3. on utilise l'estimateur de Huber $\breve{f}$ défini dans (1.3.10),
4. on ne regarde pas le point de vue minimax dans ce paragraphe, mais plutôt la vitesse du risque maximal (Définition 6) de notre estimateur de Huber.

L'adaptation dans la régression additive a déjà fait l'objet d'un travail. En effet, Brown, Cai, et Zhou [2008] utilisent la normalité asymptotique de la médiane pour approximer ce modèle par le modèle gaussien avec une méthode de médiane par blocs. Une étape intermédiaire est de projeter les nouvelles observations dans une base d'ondelettes. Ensuite, la méthode de Stein par blocs (voir Stein [1981] et Cai [1999]) est utilisée pour l'adaptation. En revanche, cette approche nécessite des hypothèses plus fortes que les hypothèses 1 qui sont suffisantes pour l'utilisation de l'estimateur de Huber.

Tsybakov [1982a, 1982b, 1983, 1986] fût le premier à introduire l'idée de Huber [1981] pour l'estimation non-paramétrique locale et plus généralement pour l'estimation robuste.

2.2.1 Adaptation

Grâce au critère de Huber, on démontre que notre estimateur $\breve{f}$ est $\phi_{n,2}(\beta)$-adaptatif sur la collection des espaces de Hölder isotropes $\left\{\mathbb{H}_d(\beta, L, M)\right\}_{\beta,L}$ pour l'estimation ponctuelle.

Pour l'estimateur de Huber, citons les papiers de Hall et Jones [1990] et Härdle et Tsybakov [1992], avec les méthodes adaptatives de *validation croisée* et de *Plug-in*. Reiss, Rozenholc, et Cuenod [2009] développe un estimateur adaptatif pour les fonction localement constantes. Dans la suite, nous proposons une généralisation de cette approche à toutes fonctions qui peuvent être, localement, approchées par un objet paramétrique.

Nous supposons que $\beta \in]0, b]$ (b est la borne supérieure de la régularité) et aucune restriction sur les constantes L et M. Soit Φ la famille de normalisations suivantes :

$$\phi_n(\beta) = \left(\frac{\rho_n(\beta)}{n}\right)^{\frac{\beta}{2\beta+d}}, \quad \rho_n(\beta) = \left(1 + \frac{2(b-\beta)}{(2\beta+d)(2b+d)} \ln n\right)^{1/2}, \quad \beta \in (0, b].$$

On remarque que $\phi_n(b) = \varphi_n(b)$ et $\rho_n(\beta) \sim \big((b-\beta)\ln n\big)^{1/2}$ pour tout $\beta \neq b$. Il est facile de démontrer que cette famille de normalisations est adaptative optimale sous le critère de Klutchnikoff [2005] pour la régression gaussienne.

Construction de l'estimateur Φ-adaptatif. Premièrement, nous déterminons la famille des *estimateurs de Huber locaux*. Ensuite, basée sur la *méthode de Lepski*, nous proposons une règle de sélection dans cette famille à partir des données.

On prend $\breve{f}^h$ l'estimateur défini par (1.3.4), (1.3.9) et (1.3.10). Ainsi, on définit la famille des estimateurs de Huber locaux $\breve{\mathcal{F}}$. Posons $h_{\max} = n^{-\frac{1}{2b+d}}$ et

$$h_k = 2^{-k} h_{\max}, \quad k = 0, \dots, \mathbf{k}_n,$$

où $\mathbf{k}_n$ est le plus grand entier tel que $h_{\mathbf{k}_n} \geq h_{\min} = n^{-1/d} \ln^{\frac{b}{d(2b+d)}} n$. Soit

$$\breve{\mathcal{F}} = \left\{ \breve{f}^{(k)}(y) = \breve{\theta}_{0,\dots,0}(h_k), \quad k = 0, \dots, \mathbf{k}_n \right\}. \tag{2.2.1}$$

On pose $\breve{f}^*(y) = \breve{f}^{(\hat{k})}(y)$, où $\breve{f}^{(\hat{k})}(y)$ est sélectionné à partir de $\breve{\mathcal{F}}$ avec la règle de sélection suivante :

$$\hat{k} = \inf\left\{ k = \overline{0, \mathbf{k}_n} : \ \left|\breve{f}^{(k)}(y) - \breve{f}^{(l)}(y)\right| \leq C S_n\big(l\big), \ \ l = \overline{k+1, \mathbf{k}_n} \right\}. \tag{2.2.2}$$

Avec les notations suivantes :

$$C = 2\, D_b \sqrt{48 \lambda r d}, \quad S_n(l) = \left[\frac{1 + l \ln 2}{n\big(h_l\big)^d}\right]^{1/2}, \quad l = 0, 1, \dots, \mathbf{k}_n, \tag{2.2.3}$$

où $\lambda > 0$ est la plus petite des valeurs propres de la matrice du design définie dans le lemme 13 et D_b est le nombre de dérivées partielles d'ordre plus petit ou égal à b (1.3.2).

Théorème 3. *Soit $b > 0$. Alors, pour toute densité $g_\xi(.)$ vérifiant les hypothèses 1, pour tout $\beta \in (0, b]$, $L > 0$, $M > 0$ et $r \geq 1$*

$$\limsup_{n \to \infty} \phi_{n,2}^{-r}(\beta) R_{n,r}\Big[\breve{f}^*(y), \mathbb{H}_d(\beta, L, M)\Big] < \infty.$$

Remarque 6. *Ce théorème correspond au théorème 5 dans le chapitre 3. La constante M sert à la construction de l'ensemble des coefficients $\Theta(M)$ (1.3.10), on peut remplacer M par $\ln n$ ou l'estimer avec une méthode de médiane locale. La constante λ dépend de $\int_{-1}^{1} g_\xi(z)dz$ (voir Preuve du lemme 13), inconnue en pratique, mais on peut l'estimer à l'aide de l'estimateur*

$$\hat{G} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[-1,1]}(Y_i - \tilde{f}(X_i)),$$

où $\tilde{f}$ est un estimateur de f (par exemple l'estimateur de la médiane locale). Il est facile de voir par construction que $\hat{G}$ estime la quantité $\mathbb{P}_f\left(-1 \le \xi \le 1\right) = \int_{-1}^{1} g_\xi(z)dz$. Une solution, très utilisée en pratique, est de calibrer la constante C avec un jeu de données.

Ce résultat est valable dans le modèle de régression additive sous les hypothèses 1, en particulier l'estimateur de Huber adaptatif $\breve{f}^*(y)$ atteint la vitesse $\phi_{n,2}(\beta)$ pour les régressions gaussienne, de Cauchy, régression α et additive uniforme.

Comme pour l'approche bayésienne, nous donnons le résultat sous forme asymptotique. Nous pouvons une nouvelle fois préciser que la borne supérieure peut s'écrire comme la vitesse plus un reste (voir Section 5.4).

Maintenant, nous allons nous intéresser à contrôler les grandes déviations de l'estimateur de Huber.

2.2.2 Grandes Déviations

Nous expliquons ici comment obtenir le contrôle des grandes déviations avec l'estimateur de Huber. Les idées de la preuve reposent sur la différentielle du critère de Huber en t, du processus limite de celle-ci (**M**-estimation) et d'un argument de chaînage combiné avec l'inégalité de Bernstein (Boucheron, Bousquet, et Lugosi [2004]). Tout d'abord, énonçons le résultat.

Pour tout $n \in \mathbb{N}^*$, $h \in \mathcal{H}_n$, $f \in \mathbb{H}_d(\beta, L, M)$, alors $\forall \varepsilon \ge D_b\, \varepsilon_0(h)/\sqrt{\lambda}$ on a

$$\mathbb{P}_f\left(N_h \left|\breve{f}^h(y) - f(y)\right| \ge \varepsilon\right) \le \Sigma \exp\left\{-\frac{\left(\varepsilon\sqrt{\lambda}D_b^{-1} - \varepsilon_0(h)\right)^2}{8 + \frac{8}{3\sqrt{nh^d}}\varepsilon\sqrt{\lambda}D_b^{-1}}\right\},$$

où $\varepsilon_0(h) = \left(1 \vee b_h\, N_h\right)$. Notons que ce résultat est conditionné par un événement dont son complémentaire a une probabilité exponentiellement petite de se réaliser (pour plus de précisions, voir Section 5.4).

Idée de la preuve. Posons la dérivée de la fonction de Huber (1.3.8)

$$q(z) := Q'(z) = \mathbb{I}_{]1,+\infty[}(z) - \mathbb{I}_{]-\infty,-1[}(z) + z\,\mathbb{I}_{[-1,1]}(z), \quad z \in \mathbb{R}. \tag{2.2.4}$$

Il est facile de voir que cette fonction est continue et bornée par 1. Notons $\tilde{D}_h(.)$ la différentielle du critère de Huber $\tilde{m}_h(.)$ défini par (1.3.9). Remarquons que chaque coordonnée de $\tilde{D}_h(.)$ est une somme de variables aléatoires bornées et $\tilde{D}_h\big(\breve{\theta}(h)\big) = (0, ..., 0)$ (q' est continue). Soit $D_h(.) = \mathbb{E}_f \tilde{D}_h(.)$, l'idée principale de cette preuve repose sur le fait que $\tilde{D}_h(t) \xrightarrow[n\to\infty]{\Delta_h + N_h^{-1}} D_h(t)$ en probabilité uniformément en t, alors on a $\breve{\theta}(h) \xrightarrow[n\to\infty]{\Delta_h + N_h^{-1}} \theta$ où $\theta \in \Theta(M)$ est l'unique solution de $D_h(\theta) = 0$ et Δ_h est le terme de biais est vaut environ $Lh\beta$. C'est l'idée générale de la m-estimation : regarder si le processus limite admet comme solution le paramètre que l'on cherche à estimer (voir par exemple le livre de Van de Geer [2000]). Pour obtenir la bonne vitesse de convergence, il suffit d'utiliser les shémas classiques du compromis du biais $Lh\beta$ et de la variance $1/sqrtnh^d$. On donne le diagramme suivant qui résume ces explications (Figure 2.1).

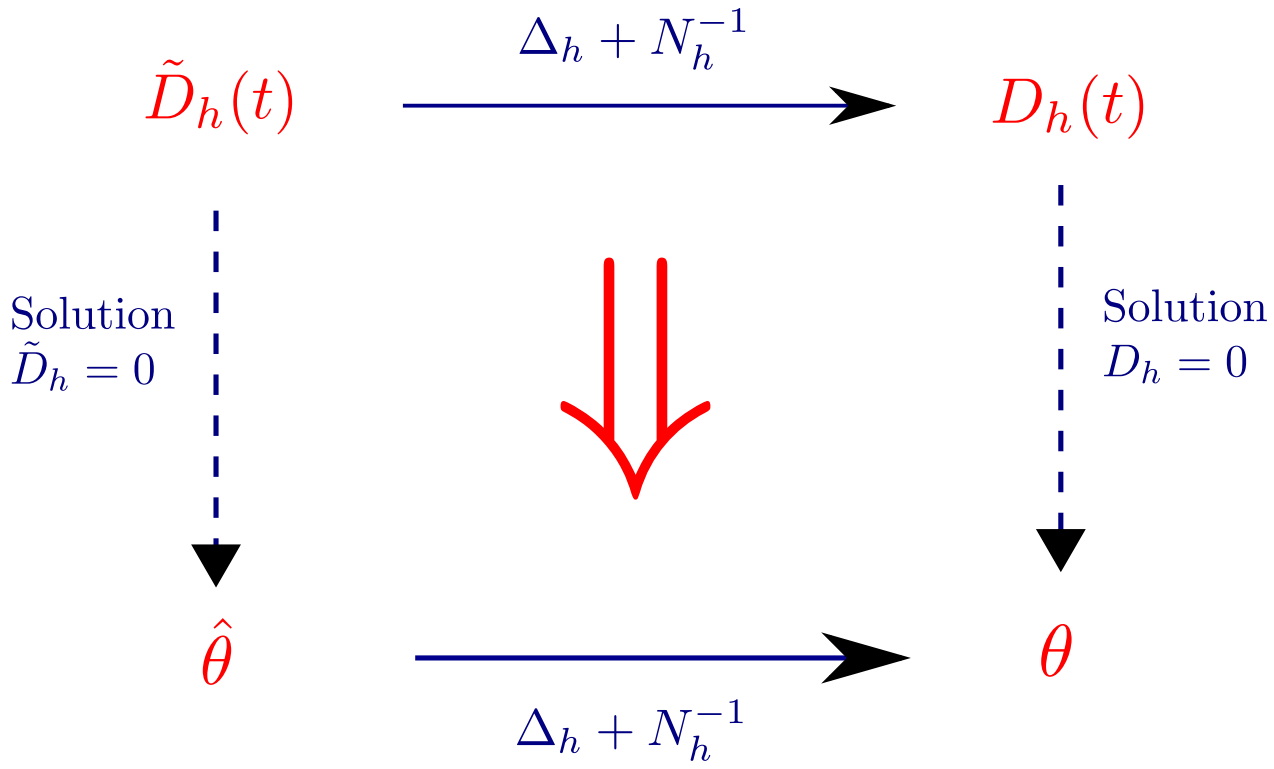


FIGURE 2.1 – Illustration des principe des M-estimateurs

La définition de $\breve{\theta}(h)$ et $\theta = \theta(f, y, h)$ les coefficients du développement de Taylor impliquent que $\forall \varepsilon \geq D_b\, \varepsilon_0(h)/\sqrt{\lambda}$

$$\begin{aligned}\mathbb{P}_f\left(N_h\big|\breve{f}^h(y) - f(y)\big| \geq \varepsilon\right) &\leq \mathbb{P}_f\left(N_h\big|\breve{\theta}_0(h) - \theta_0\big| \geq \varepsilon\right)\\ &\leq \mathbb{P}_f\left(N_h\sqrt{D_b}\big\|\breve{\theta}(h) - \theta\big\|_2 \geq \varepsilon\right).\end{aligned}$$

En démontrant que la jacobienne de $D_h(.)$ est inversible, on peut écrire (Lemme 13) que

$$\mathbb{P}_f\left(N_h\big|\breve{f}^h(y)-f(y)\big|\geq\varepsilon\right)\leq\mathbb{P}_f\left(N_h\frac{\sqrt{D_b}}{\sqrt{\lambda}}\Big\|D_h\big(\breve{\theta}(h)\big)-D_h(\theta)\Big\|_2\geq\varepsilon\right).$$

Maintenant, rappelons que $\tilde{D}_h\big(\breve{\theta}(h)\big)=(0,...,0)$ et $D_h(\theta)=(0,...,0)$ et

$$\Big\|D_h\big(\breve{\theta}(h)\big)-D_h(\theta)\Big\|_2=\Big\|D_h\big(\breve{\theta}(h)\big)\Big\|_2=\Big\|D_h\big(\breve{\theta}(h)\big)-\tilde{D}_h\big(\breve{\theta}(h)\big)\Big\|_2\leq\sup_{t\in\Theta(M)}\Big\|D_h\big(t\big)-\tilde{D}_h\big(t\big)\Big\|_2.$$

Ainsi

$$\mathbb{P}_f\left(N_h\big|\breve{f}^h(y)-f(y)\big|\geq\varepsilon\right)\quad\leq\quad\mathbb{P}_f\left(N_h\frac{\sqrt{D_b}}{\sqrt{\lambda}}\sup_{t\in\Theta(M)}\Big\|D_h\big(t\big)-\tilde{D}_h\big(t\big)\Big\|_2\geq\varepsilon\right)$$

En utilisant un argument de chaînage et l'inégalité de Bernstein (voir Lemme 15), on obtient le résultat

$$\mathbb{P}_f\left(N_h\big|\breve{f}^h(y)-f(y)\big|\geq\varepsilon\right)\leq\Sigma\exp\left\{-\frac{\left(\varepsilon\sqrt{\lambda}D_b^{-1}-\Delta_h\right)^2}{8+\frac{8}{3\sqrt{nh^d}}\varepsilon\sqrt{\lambda}D_b^{-1}}\right\}. \tag{2.2.5}$$

Pour plus de détails, regarder la proposition 7 dans la section 5.4. On peut voir $D_h\big(t\big)-\tilde{D}_h\big(t\big)$ comme une somme de variable aléatoires indépendantes et bornées $\sum_{i=1}^n\mathcal{W}_t(X_i,\xi_i)$ (processus empirique). Deux nombreux outils ont été développés ces quinze dernières années et peuvent remplacer l'utilisation d'un argument de chaînage et d'une inégalité de concentration classique.

2.2.3 Inégalités Maximales pour les processus empiriques

Cette sous-section a pour but de présenter la notion de *processus empirique* et les inégalités de concentration du maxima de ces processus que l'on peut trouver dans la littérature. Ces inégalités sont très utilisées en statistique mathématique notamment pour l'adaptation (Voir l'exemple ci-dessus). Elles sont couramment appelées *inégalités maximales* (en anglais : *maximal inequalities*).

Posons les notations pour définir un processus empirique. Soit $\mathcal{G}$ une classe de fonctions bornées de $\Omega\to\mathbb{R}$ où $(\Omega,\mathcal{A},\mathbb{P})$ est un espace de probabilité et $\mathbb{E}$ désigne l'espérance mathématique par rapport à $\mathbb{P}$. Notons $Z_1,...,Z_n$ une suite de variables aléatoires indépendantes à valeurs dans Ω. On appelle *processus empirique* $\sum_{i=1}^n g(Z_i)$ et maxima d'un processus empirique $Z=\sup_{g\in\mathcal{G}}\sum_{i=1}^n\big|g(Z_i)-\mathbb{E}g(Z_i)\big|$.

Inégalité de Talagrand :

Pour la première fois, Talagrand [1994, 1995, 1996a, 1996b] donne des inégalités de concentration pour le maxima d'un processus empirique. Soient $b > 0$ tel que $\|g\|_\infty \leq b$ pour tout $g \in \mathcal{G}$ et $v = \mathbb{E}\sup_{g\in\mathcal{G}} \sum_{i=1}^n g^2(Z_i)$, alors pour tout réel positif

$$\mathbb{P}\left(Z \geq \mathbb{E}[Z] + x\right) \leq K \exp\left\{-\frac{x^2}{2(c_1 v + c_2 b x)}\right\},$$

où c_1, c_2 et K sont des constantes universelles.

On peut remarquer que si $\mathcal{G}$ est restreint à une seule fonction, alors on obtient *l'inégalité de Bernstein*

$$\mathbb{P}\left(\sum_{i=1}^n \left|g(Z_i) - \mathbb{E}g(Z_i)\right| \geq x\right) \leq 2\exp\left\{-\frac{x^2}{2(v + bx/3)}\right\}.$$

Dans ce cas, les constantes correspondantes c_1 et c_2 sont connues explicitement. L'inégalité de Talagrand permet de prouver l'existence d'inégalités exponentielles, mais pose le problème de la connaissance des constantes. En effet, les procédures adaptatives contiennent des paramètres qui, une fois optimisés, dépendent des constantes c_1 et c_2 (Voir le paramètre C dans la procédure de Lepski). De plus, il n'est pas toujours évident de calculer l'espérance de $\mathbb{E}[Z]$ (pour plus de détails, voir Massart [2007], Chapitre 6). En fait, cela demande l'utilisation d'un argument de chaînage et donc du calcul de l'entropie de l'espace $\mathcal{G}$ considéré.

Inégalité de Massart 1 :

Pour résoudre les problèmes que pose l'inégalité de Talagrand, Ledoux [1997] propose une démonstration de cette inégalité, en utilisant des inégalités d'entropie sur des mesures produit. Puis Massart [2000] donne une inégalité de type Bennett avec des constantes connues. On peut toujours écrire une inégalité de type Bennett comme une inégalité de type Bernstein qui est souvent privilièGée pour l'adaptation. Soit $\sigma^2 = \sup_{g\in\mathcal{G}} \sum_{i=1}^n Var\big(g(Z_i)\big)$, alors pour tous réels strictement positifs z et ε

$$\mathbb{P}\left(Z \geq (1+\varepsilon)\mathbb{E}[Z] + x\right) \leq \exp\left\{-\frac{x^2}{2(\kappa\sigma^2 + \kappa(\varepsilon) b x)}\right\},$$

où $\kappa = 4$ et $\kappa(\varepsilon) = 2.5 + 32\varepsilon^{-1}$.

La connaissance des constantes est très importante pour l'adaptation. Le problème, qui vient ensuite, est l'optimisation de ces constantes (les plus petites possibles).

Inégalité de Bousquet :

Contrairement à l'inégalité de Massart 1, Bousquet [2002] montre que l'on peut obtenir les constantes optimales (comme dans l'inégalité de Bernstein) avec des variables aléatoires $Z_1, ..., Z_n$ indépendantes et identiquement distribuées. Supposons que $\sup_{g\in\mathcal{G}} \|g\|_\infty \leq 1$, alors pour tout $x \geq 0$

$$\mathbb{P}\left(Z \geq \mathbb{E}[Z] + x\right) \leq \exp\left\{-\frac{x^2}{2(\sigma^2 + x/3)}\right\}.$$

Même avec les constantes optimales, il reste un problème de taille à résoudre : le calcul ou la majoration de $\mathbb{E}[Z]$ (l'espérance du maxima de la somme de variables aléatoires indépendantes et bornées). On peut utiliser un résultat disponible dans le livre de Van der Vaart et Wellner [1996] pour les processus $(\mathcal{X}_t)_{t\in T}$ dit sub-gaussiens quand ils vérifient

$$\mathbb{P}\left(|\mathcal{X}_s - \mathcal{X}_t| > x\right) \leq 2\exp\left\{-x^2/2d^2(s,t)\right\},\ s,t\in T,\ x>0,$$

où $d(\cdot,\cdot)$ est une semi-métrique sur T. Dans le chapitre 2 (*Empirical processes*), Section 2 (*Maximal inequalies*), corollaire 2.2.8, on y trouve une majoration de l'espérance du maxima du processus $\mathcal{X}_t$: pour $t_0\in T$ fixé arbitrairement, on a

$$\mathbb{E}\sup_{t\in T}|\mathcal{X}_t| \leq \mathbb{E}|\mathcal{X}_{t_0}| + K\int_0^\infty \sqrt{\ln N(r/2,d)}dr,$$

pour K une constante universelle et $N(r,d)$ est le nombre minimal de boules de rayon r (avec la semi-métrique d) nécessaire pour recouvrir l'espace T. $\ln N(r/2,d)$ est appelée *entropie* de l'espace T.

La démonstration de ce corollaire nécessite l'utilisation d'un argument de chaînage et donc le calcul de l'entropie de l'espace considéré. Mais ici, la constante K est inconnue. Ceci pose encore le problème de l'adaptation dans lequel les paramètres des procédures adaptatives doivent être connues. Dans ce sens, Massart [2007] développa des inégalités maximales qui nécessitent seulement le calcul de l'entropie.

Inégalité de Massart 2 :

On peut trouver dans le livre de Massart [2007] au Chapitre 6 (*Maximal Inequalities*), Section 6.2 (*Function-indexed empirical processes*), des inégalités exponentielles pour le maxima de processus empiriques (somme de variables aléatoires indépendantes et bornées). Nous expliquerons dans la suite que si le supremum du processus empirique $\sum_i \mathcal{W}_t(X_i,\xi_i)$ est indéxé en t alors il est nécessaire que $\mathcal{W}_t(\cdot,\cdot)$ soit continue en t. Ainsi, le résultat énoncé par Massart [2007] dans le corollaire 6.9 se présente comme suit. L'auteur introduit la fonction $H(\delta)$ comme l'entropie de $\mathcal{G}$ avec une notion de *Bracketing* (un peu plus rafinée que l'entropie habituelle). Soit $\sigma^2 = \sup_{g\in\mathcal{G}}\sum_{i=1}^n \mathbb{E}g^2(Z_i)$ et $b = \sup_{g\in\mathcal{G}}\|g\|_\infty$, alors, pour tout $\varepsilon\in]0,1]$ et $x\geq 0$

$$\mathbb{P}\left(Z\geq E+x\right) \leq \exp\left\{-\frac{x^2}{2(1+6\varepsilon)^2 n\sigma^2 + 4bx/3}\right\},$$

où pour une constante absolue $\kappa = 27$

$$E = \frac{\kappa}{\varepsilon}\sqrt{n}\int_0^{\varepsilon\sigma}\sqrt{H(u)\vee n}du + 2(b/3+\sigma)H(\sigma)$$

Ici tout le travail a été fait, en particulier la majoration de $\mathbb{E}Z$ par E qui contient l'intégrale de l'entropie avec les contantes connues.

Revenons au cas de l'estimateur de Huber, avec le maxima du processus empirique $\sup_{t\in\Theta}\sum_i \mathcal{W}_t(X_i,\xi_i)$ ($\Theta=\Theta(M)$ est un compact de $\mathbb{R}^{D_b}$) où $\mathcal{W}_t(x,z)=q\big(z+f(x)-f_t(x)\big)\left(\frac{x-y}{h}\right)\mathbb{I}_{x\in V_h(y)}$, et q est la dérivée de la fonction de Huber definie dans (1.3.8). Comme nous l'avons précisé ci-dessus, le calcul de l'entropie sur l'espace $\mathcal{G}=\mathbb{W}_\Theta=\big\{\mathcal{W}=\mathcal{W}_t\,:\,t\in\Theta\big\}$ est assez difficile voir impossible. Ainsi, il est souvent choisi de se ramener à l'entropie de l'espace Θ qui est plus facile à calculer. Ceci nécessite la continuité de $\mathcal{W}_t$ en t et donc de la continuité de la fonction q (dérivée de Huber) dans notre exemple à nous.

L'inégalité de Massart 2 convient parfaitement pour le contrôle des grandes déviations de l'estimateur de Huber. Mais dans notre démonstration (Chapitre 5, Preuve de la proposition 7), nous avons choisi de faire les calculs à la *main* avec l'utilisation d'un argument de chaînage et l'inégalité de Bernstein (l'inégalité de Hoeffding est suffisante si le design est déterministe).

2.3 Expériences numériques

Dans cette section, nous donnons quelques résultats pratiques (avec le logiciel *Matlab*) pour toutes les méthodes développées. Nos simulations sont effectuées sur 3 fonctions f_1, f_2 et f_3.

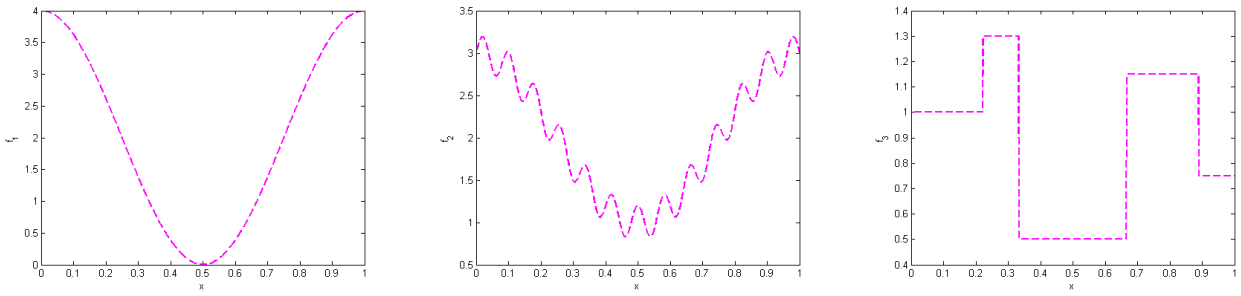


FIGURE 2.2 – Fonctions utilisées pour nos simulations.

Notre étude se limite aux fonctions unidimensionnelles. En fait, les comportements, que l'on peut observer sur les fonctions isotropes à plusieurs variables, sont semblables à ceux observés sur les trois fonctions ci-dessus.

Rappelons que notre estimateur est un polynôme dont on estime les coefficients. Pour des raisons de complexité, nous nous limiterons aux polynômes de degré 1, i.e. que nous n'étudierons que les fonctions $f \in \mathbb{H}_d(\beta, L, M)$ avec $\beta \leq 2$.

Nous testons nos estimateurs dans les différents modèles étudiés dans cette thèse. Pour chacun d'entre eux, nous allons comparer les estimateurs bayésiens et de Huber avec leurs oracles.

Définition 10. *L'estimateur* $\mathring{f}^*(y) = \mathring{f}^{h^*}(y)$ *est appelé oracle de la famille* $\left\{\mathring{f}^h(y)\right\}_{h>0}$ *si*

$$h^* = \arg\min_{h>0} \mathbb{E}_f \big|\mathring{f}_h(y) - f(y)\big|.$$

h^* *est appelée fenêtre oracle.*

Notons que pour chaque point y, on y associe un oracle. L'oracle peut être interprété comme le meilleur estimateur dans la famille connaissant la fonction f.

Nous présentons les résultats de la façon suivante : nous calculons les rapports

$$r_B^*(y) = \frac{\mathbb{E}_f\big|\hat{f}_B^*(y) - f(y)\big|}{\mathbb{E}_f\big|\hat{f}_B^{h_B^+}(y) - f(y)\big|}, \qquad r_H^*(y) = \frac{\mathbb{E}_f\big|\breve{f}_H^*(y) - f(y)\big|}{\mathbb{E}_f\big|\breve{f}_H^{h_H^+}(y) - f(y)\big|},$$

où $\hat{f}_B^*(y)$ et $\breve{f}_H^*(y)$ sont respectivement les oracles des familles d'estimateurs bayésiens $\left\{\hat{f}^h(y)\right\}_{h>0}$ et de Huber $\left\{\breve{f}^h(y)\right\}_{h>0}$. h_B^+ et h_H^+ sont respectivement les fenêtres adaptatives des estimateurs bayésien et de Huber, données par les procédures (2.1.5) et (2.2.2). L'espérance de la

fonction de perte est approchée par une méthode de Monte-Carlo avec $T = 10000$ itérations. On construit la suite finie d'éléments $y_1, y_2, ..., y_v$ uniformément répartis sur l'intervalle $[0, 1]$. On appelle *oracle bayésien / estimateur bayésien* et *oracle de Huber / estimateur de Huber* les rapports suivants :

$$r_B^* = \frac{1}{v}\sum_{j=1}^{v} r_B^*(y_j), \qquad r_H^* = \frac{1}{v}\sum_{j=1}^{v} r_H^*(y_j).$$

Nous choisirons $v = 200$ pour l'implémentation. Ces rapports permettent d'évaluer la qualité des procédures adaptatives que nous avons introduit dans cette thèse (voir Sections 2.1 et 2.2). Rappelons que ces procédures contiennent un paramètre de seuillage, noté C (voir (2.2.3) et (2.1.6)). Nous proposons un choix théorique connu de ce paramètre, mais en pratique, on peut le choisir de façon optimale. La figure 2.3 illustre le choix optimal de ce paramètre. En effet, il suffit de le calibrer par rapport à un jeu de données ou de fonctions *fantômes* pour calculer l'oracle et regarder la distance qui sépare notre estimateur à cet oracle.

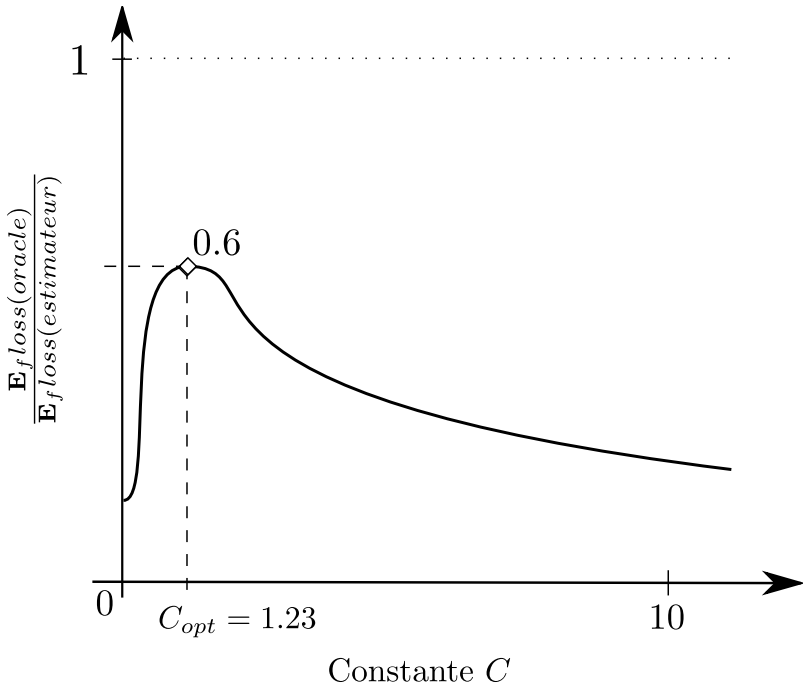


FIGURE 2.3 – Illustration de la calibration de la constante C à travers un jeu de données connu. Ici, le modèle est celui de la régression gaussienne (avec la fonction f_1), avec $n = 1000$ observations. Nous itérons $T = 10000$ séries d'observations pour approximer l'espérance du risque. La fenêtre de l'estimateur bayésien adaptatif est choisie avec la procédure donnée dans la section 2.1. Nous avons remarquer que le C_{th} est cinq à dix fois plus grand que le C_{opt}.

Nous constatons ici une nette différence entre le risque de notre estimateur et celui de l'oracle (avec un rapport légérement supérieur à 1/2). Ceci met en évidence les points faibles de l'approche adaptative au sens minimax et de l'estimation ponctuelle (prix à payer). En effet, Nous aimerions que notre rapport r^* soit le plus proche possible de 1, au minimum en C_{opt}. Une perspective serait d'utiliser une méthode adaptative pour obtenir des inégalités oracle (Voir Section 2.4).

Nous donnons aussi, quand cela est possible, les rapports *oracle linéaire / estimateur bayésien* et *oracle linéaire / estimateur de Huber*, notés r^*_{LB} et r^*_{LH}, où l'oracle Linéaire est celui d'une famille d'estimateurs à noyau classique (noyau d'Epanechnikov). Cela nous permet de comparer la qualité des nouveaux estimateurs, développés dans cette thèse, avec les plus classiques comme les estimateurs linéaires.

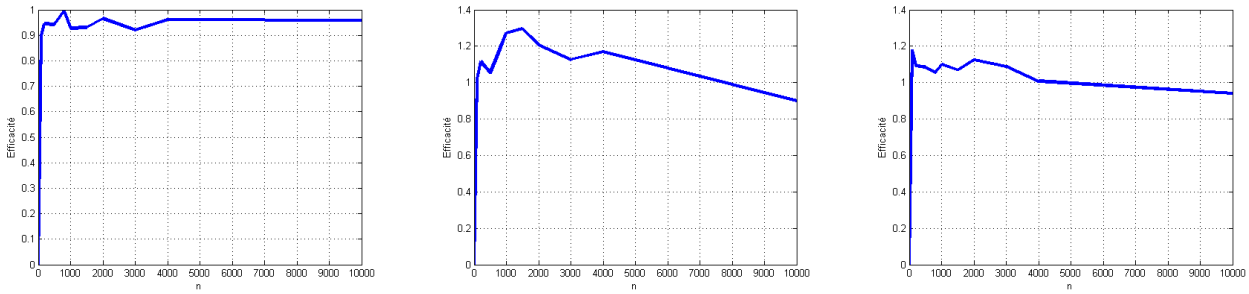
Rappelons que la complexité des estimateurs bayésien et de Huber est plus grande que celle des estimateurs linéaires. Pour le critère de Huber qui est convexe, cela peut se faire avec une méthode de descendte du gradient. En revenche, le critère bayésien n'est pas du tout évident à minimiser surtout si la densité des observations est discontinue. Ceci est le point faible de ces nouveux estimateurs (non explicites). En revanche, avec le même d'observations, l'estimateur bayésien est plus rapide dans certains cas, ce qui entraîne un besoin moins important en observations. En pratique, la construction d'observations peut coûter cher en temps et en argent. Un autre aspect est de considérer les modèles où les estimateurs linéaires ne sont pas consistants (par exemple, modèle de Cauchy). Ainsi, nous pouvons utiliser l'estimateur de Huber pour construire un estimateur au moins consistant.

Légende. Pour chaque modèle, nous illustrons nos simulations en traçant les courbes r^*_B, r^*_H, r^*_{LB} et r^*_{LH} en fonction de n (jusqu'à $n = 10000$). Ceci est effectué pour chaque fonction de régression f_1, f_2 et f_3. Dans le même temps, nous calculons la probabilité $\mathbb{P}_f\big(h^* < h^+\big)$ pour chaque estimateur adaptatif. Cette probabilité correspond aux grandes déviations de l'estimateur en question. Le choix du paramètre de seuillage C est fait de façon optimale (voir Figure 2.3) et est donné pour chaque modèle.

Régression Gaussienne

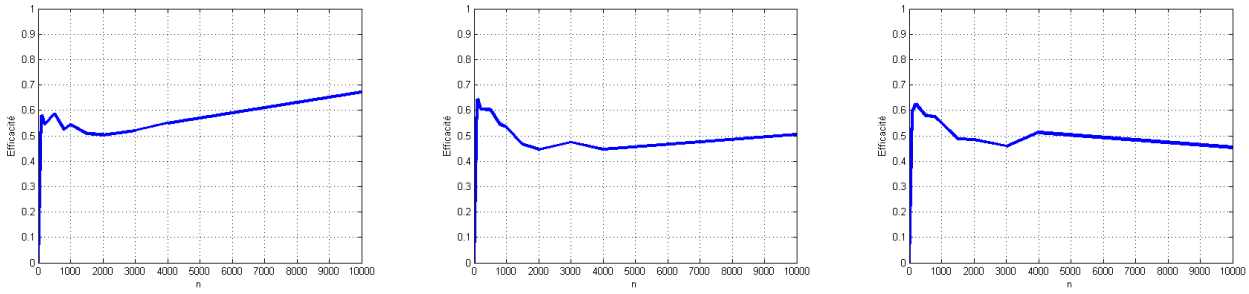
On peut voir sur la figure 2.4 que l'oracle bayésien fait aussi "bien" que l'oracle linéaire. Ce qui confirme les résultats théoriques obtenus pour ce modèle (voir Section 2.1.4). A une constante près, les deux estimateurs ont la même vitesse $\phi_{n,2}$. La famille d'estimateurs bayésiens est donc aussi "riche" que celle des estimateurs linéaires.

On constate sur cette illustration (Figure 2.5) que le risque de l'estimateur adaptatif est deux fois plus grand que celui de l'oracle. Ceci est classique en estimation ponctuelle du fait du *prix à payer* ($\ln^{1/\gamma} n$) pour l'adaptation (Voir Théorème 2). Il faudrait prendre un nombre d'observations $n \gg 10000$ (bien plus grand que 10000) pour voir le rapport tendre vers 1. Ce phénomène s'explique en particulier par le fait que les grandes déviations sont environ égales à 4%.



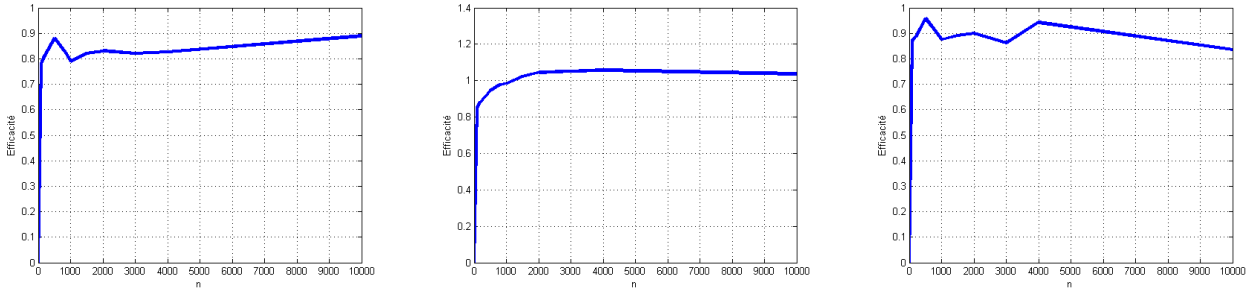
r^*_{LB,f_1} r^*_{LB,f_2} r^*_{LB,f_3}

FIGURE 2.4 – Rapport *oracle linéaire / oracle bayésien* (modèle gaussien).



r^*_{B,f_1} r^*_{B,f_2} r^*_{B,f_3}

FIGURE 2.5 – Rapport *oracle bayésien / bayésien adaptatif* : $C_{opt} = 1.2$ and $\mathbb{P}\big(\hat{h} < h^*\big) \simeq 0.0411$ (modèle gaussien).

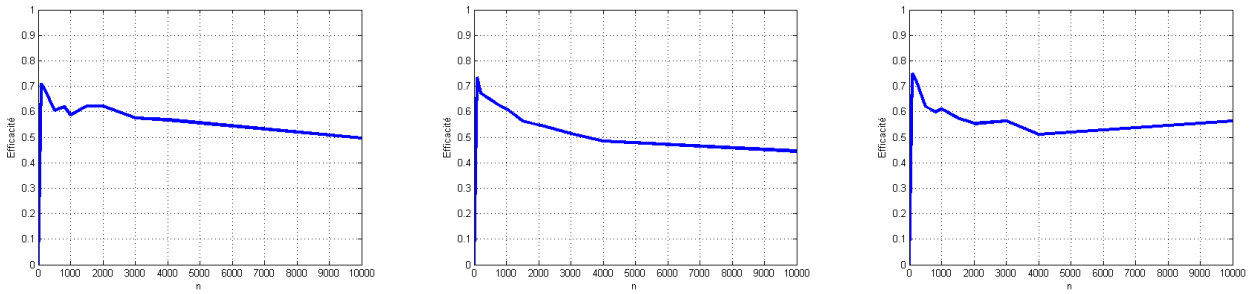


r^*_{LH,f_1} r^*_{LH,f_2} r^*_{LH,f_3}

FIGURE 2.6 – Rapport *oracle Linéaire / oracle Huber* (modèle gaussien).

De même que l'oracle bayésien, l'oracle de Huber fait aussi bien que l'oracle linéaire (Figure 2.6). Nous signalons que le rapport est légèrement en-dessous de 1 (plutôt 0.9). Ceci est dû aux constants de la borne supérieure pour l'estimateur de Huber. Cette constante est en effet plus grande que celle que l'on peut obtenir pour un estimateur linéaire. Ce

phénomène se retrouve entre la médiane et la moyenne empirique.

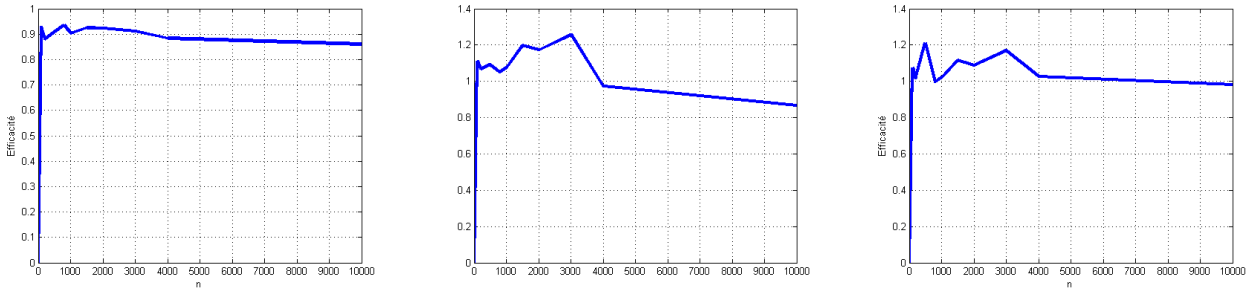


r^*_{H,f_1} r^*_{H,f_2} r^*_{H,f_3}

FIGURE 2.7 – Rapport *oracle Huber / Huber adaptatif* : $C_{opt} = 1.2$ and $\mathbb{P}\big(\hat{h} < h^*\big) \simeq 0.065$ (modèle gaussien).

Nous constatons le même effet que dans le cas précédent. L'estimateur adaptatif de Huber est deux fois moins bon que l'oracle de Huber (Figure 2.7). Comme cette remarque a déjà été faite pour l'estimateur bayésien adaptatif, nous pensons que cela est dû au modèle avec des grandes déviations égales à 6.5%. La forme de l'estimateur en question ne joue aucun rôle, nous mettrons en avant la sensibilité de la méthode de Lepski par rapport au modèle. Nous verrons dans la suite d'autres modèles où les valeurs des rapports *oracle / estimateur adaptatif* changent.

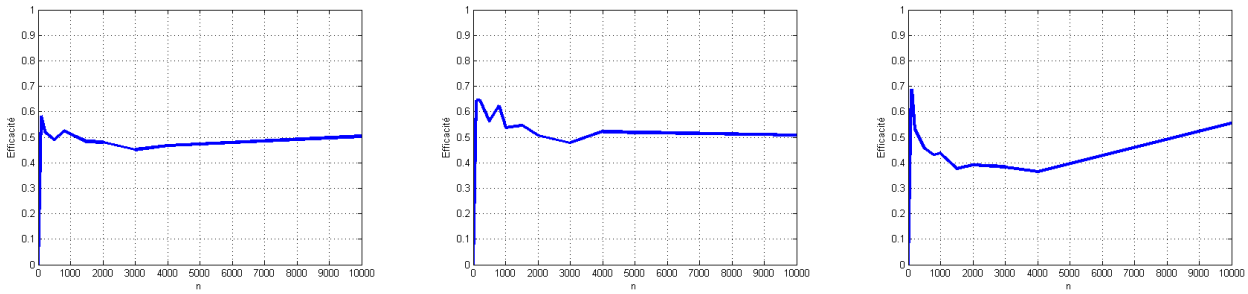
Régression Inhomogène de Poisson



r^*_{LB,f_1} r^*_{LB,f_2} r^*_{LB,f_3}

FIGURE 2.8 – Rapport *oracle linéaire / oracle bayésien* (Modèle de Poisson).

Le modèle de Poisson est asymptotiquement équivalent au modèle gaussien (voir Anscombe [1948]). De ce fait, on retrouve les mêmes caractéristiques que le modèle gaussien. L'oracle bayésien est aussi bon que l'oracle linéaire (Figure 2.8). Le rapport *oracle bayésien / bayésien adaptatif* vaut environ 1/2 avec $\mathbb{P}\big(\hat{h} < h^*\big) \simeq 0.0480$ (Figure 2.9).

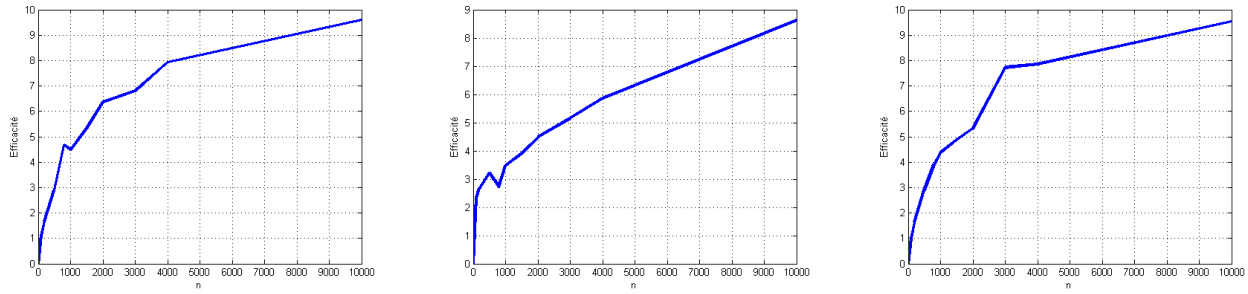


r^*_{B,f_1} r^*_{B,f_2} r^*_{B,f_3}

FIGURE 2.9 – Rapport *oracle bayésien / bayésien adaptatif* : $C_{opt} = 1.2$ and $\mathbb{P}\big(\hat{h} < h^*\big) \simeq 0.0480$ (Modèle de Poisson).

Régression $\alpha = 1/4$

Nous avons choisi pour nos simulations $\alpha = 1/4$.

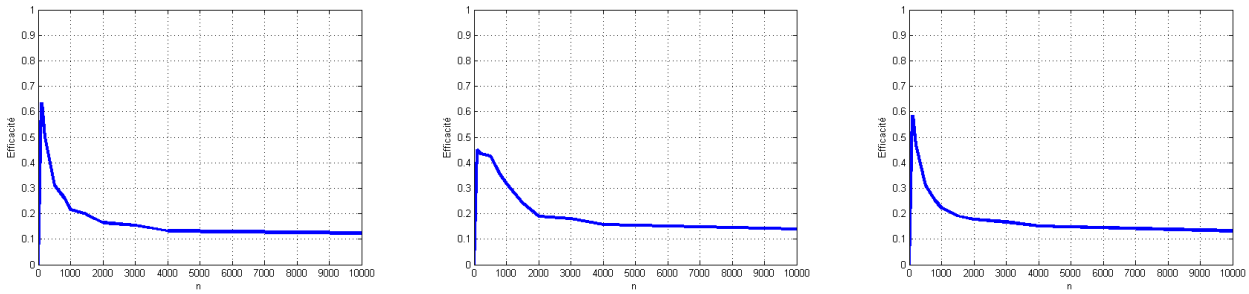


r^*_{LB,f_1} r^*_{LB,f_2} r^*_{LB,f_3}

FIGURE 2.10 – Rapport *oracle linéaire / oracle bayésien* (modèle α).

Comme nous l'avons vu dans la section 2.1.4, la vitesse atteinte par l'estimateur bayésien est $\phi_{n,1.5}$. Elle est donc meilleure que la vitesse des estimateurs linéaires $\phi_{n,2}$. Nous constatons ce phénomène sur les simulations du rapport *oracle linéaire / oracle bayésien* (Figure 2.10). Ici, le rapport peut atteindre la valeur de 10 pour $n = 10000$, i.e que l'oracle bayésien est dix fois plus rapide que l'oracle linéaire.

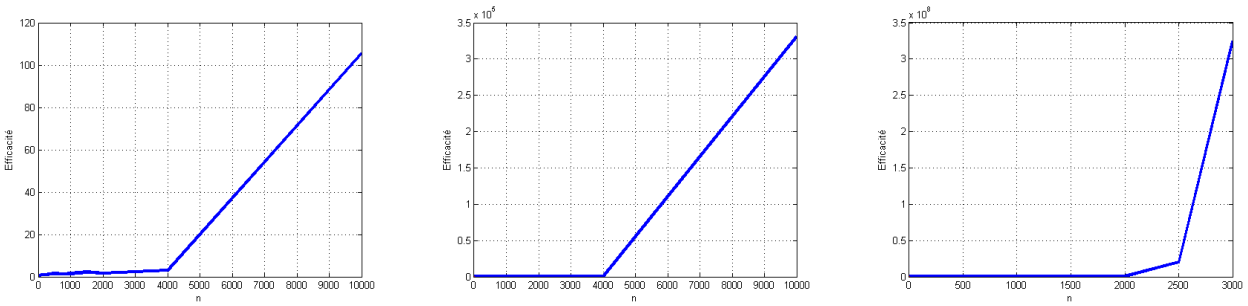
L'étude de la performance de la méthode adapatative (méthode de Lepski) pour la régression α, révèle que celle-ci est assez faible. En effet, le rapport *oracle bayésien / bayésien adaptatif* est quasiment égal à 0.1 (Figure 2.11). Ce qui veut dire que le risque de l'estimateur adapatif est 10 fois moins bon que celui de l'oracle. On l'explique par le fait que les grandes déviations sont relativement élevées (par rapport aux modèles précédents, égales à 17%). Ce modèle comporte un certain nombre de valeurs extrêmes qui peut nous laisser penser que celles-ci mettent à mal la méthode de Lepski. En effet, nous avons mentionné dans la section 1.5.2 que cette méthode s'arrête trop tôt en précence de valeurs extrêmes.



r^*_{B,f_1} r^*_{B,f_2} r^*_{B,f_3}

FIGURE 2.11 – Rapport *oracle bayésien / bayésien adaptatif* : $C_{opt} = 2$ and $\mathbb{P}\big(\hat{h} < h^*\big) \simeq 0.1703$ (modèle α).

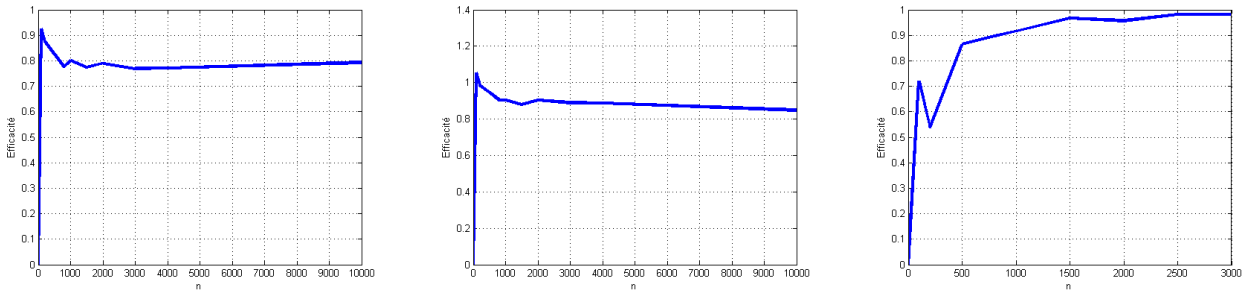
Régression Multiplicative Uniforme



r^*_{LB,f_1} r^*_{LB,f_2} r^*_{LB,f_3}

FIGURE 2.12 – Rapport *oracle linéaire / oracle bayésien* (modèle multiplicatif uniforme).

Dans cette illustration (Figure 2.12), on constate un net avantage pour l'oracle bayésien. En effet, pour la fonction f_1, le rapport *oracle linéaire / oracle bayésien* atteint 250. Ensuite, pour les fonctions irrégulières f_2 et f_3, nous avons des valeurs qui sont de l'ordre de 10^5 et 10^8. Remarquons que pour la fonction f_3, nous nous sommes arrêté à $n = 3000$, à cause des valeurs trop petites du risque de l'oracle bayésien que le logiciel *Matlab* arrondit à 0. Deux remarques permettent d'expliquer les bonnes performances de l'estimateur bayésien. Premièrement, la vitesse atteinte $\phi_{n,1}$ est meilleure que $\phi_{n,2}$. Deuxièmement, pour les fonctions irrégulières, l'estimateur bayésien est plus performant pour l'estimation au voisinage des points de discontinuité, tandis que l'estimateur linéaire devient très mauvais pour ce genre de fonctions.



r^*_{B,f_1} r^*_{B,f_2} r^*_{B,f_3}

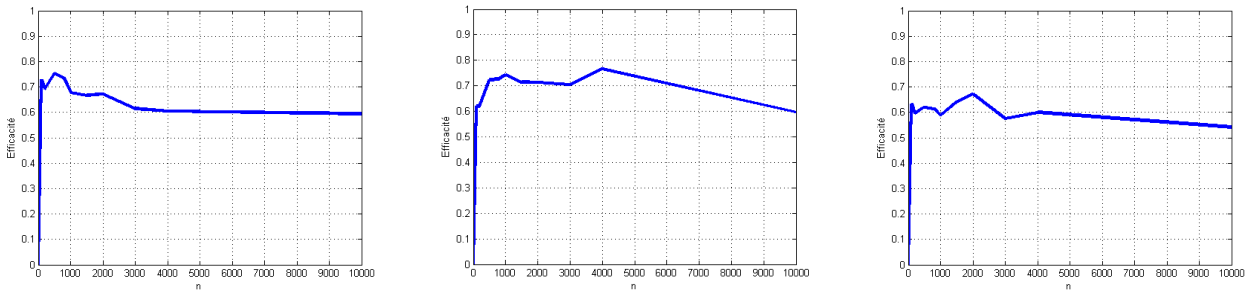
FIGURE 2.13 – Rapport *oracle bayésien / bayésien adaptatif* : $C_{opt} = 2$ and $\mathbb{P}\big(\hat{h} < h^*\big) \simeq 0$ (modèle multiplicatif uniforme).

En constatant que l'estimateur adaptatif fait aussi bien que son oracle (Figure 2.13), nous pouvons souligner une nouvelle fois la robustesse de notre estimateur, car les grandes déviations sont nulles. Ce dernier modèle souligne une nouvelle fois, l'intérêt de trouver la vitesse optimale et l'estimateur qui l'atteigne. Dans ce sens, l'estimateur bayésien joue pleinement son rôle.

Régression de Cauchy

Pour le modèle de Cauchy, il n'est pas possible de construire les estimateurs linéaires, du fait qu'ils ne sont pas consistants. Nous donnons, pour évaluer la performance de notre estimateur de Huber, le risque de l'estimateur adaptatif en moyenne pour les trois fonctions ($n = 10000$) :

$$\frac{1}{v}\sum_{j=1}^{v} \mathbb{E}_f \big| \check{f}_H^{h_H^+}(y_j) - f(y_j) \big| \simeq 10^{-1}.$$



r^*_{B,f_1} r^*_{B,f_2} r^*_{B,f_3}

FIGURE 2.14 – Rapport *oracle de Huber / Huber adaptatif* : $C_{opt} = 3$ and $\mathbb{P}\big(\hat{h} < h^*\big) \simeq 4.7\%$ (modèle de Cauchy).

Ceci est une bonne performance si on prend en compte les constantes de la borne supérieure et la vitesse qui vaut environ $1/\sqrt{10000} = 10^{-2}$ pour les fonctions discontinues.

Pour la partie adaptation, nous constatons que le rapport *oracle de Huber / Huber adaptatif* vaut 0.6 pour les trois fonctions (Figure 2.14). Nous soulignons le caractère robuste de l'estimateur de Huber qui, malgré un bruit de Cauchy (avec un grand nombre de valeurs extrêmes), rend stable la procédure de Lepski avec seulement $\mathbb{P}\big(\hat{h} < h^*\big) \simeq 0.047\%$.

2.4 Perspectives

Nous avons présenté dans cette thèse deux nouvelles méthodes d'estimation : l'une qui recherche la vitesse optimale (bayésienne) et l'autre qui est robuste et qui ne dépend pas de la densité du bruit (Huber). Nous donnons, dans la suite, une quinzaine de problèmes ouverts relatifs à ces deux approches.

2.4.1 Approche Bayésienne

1. L'approche bayésienne est déduite des travaux de Has'minskii et Ibragimov [1981] pour les modèles paramétriques. Nous avons insisté sur le fait que le processus $Z_{n,\theta}(.)$ (rapport de vraisemblance) doit vérifier les hypothèses 3. Mais en réalité, on peut prendre n'importe quel processus qui vérifie ces hypothèses. Par exemple, dans le modèle de la régression additive uniforme définie par

$$Y_i = f(X_i) + U_i, \quad U_i \sim \mathcal{U}_{[-1,1]}, \quad i = 1, ..., n.$$

   Avec ce modèle, nous ne pouvons pas définir le processus $Z_{n,\theta}(.)$ car les mesures de probabilité ne sont absolument pas continues.

   Problème Ouvert 1. *Choisir un processus $Z'_{n,\theta}$ qui vérifie les hypothèses 3 et qui permet de démontrer que l'estimateur bayésien atteint la vitesse minimax $n^{-\frac{\beta}{\beta+d}}$ pour la régression additive uniforme ci-dessus. Ensuite, montrer que l'estimateur bayésien adaptatif atteint la vitesse $\phi_{n,\gamma}$.*

2. Pour le modèle multiplicatif uniforme, nous avons émis des hypothèses supplémentaires $0 < A \leq f \leq M$. La borne supérieure de la fonction f est une hypothèse classique. En revanche, l'hypothèse de positivité de celle-ci est absolument nécessaire pour l'approche bayésienne mais non justifiée.

   Problème Ouvert 2. *Pour le modèle de régression multiplicatif uniforme, affaiblir l'hypothèse de positivité de la fonction f à estimer, i.e. supposer seulement que $\|f\|_\infty < \infty$.*

3. Nous pensons que l'on peut atteindre la vitesse $n^{-\frac{\beta}{\beta+d}}$ pour tout modèle où la densité des observations admet un nombre fini de points de discontinuité. Cela a déjà été démontré dans le cas paramétrique par Has'minskii et Ibragimov [1981] (Chapitre 5, Section 1).

   Problème Ouvert 3. *Soit le modèle de régression générale (voir 1.2.2), où la densité g admet un nombre fini de points de discontinuité. Démontrer que*

la vitesse $n^{-\frac{\beta}{\beta+d}}$ est atteinte par l'estimateur bayésien, i.e. que les hypothèses 3 sont vérifiées avec $\gamma = 1$.

4. Dans cette thèse, nous avons supposé une hypothèse d'indépendance des observations, mais celle-ci n'est pas nécessaire. On pourra alors traiter des modèles avec des observations dépendantes.

   Problème Ouvert 4. *Utiliser l'approche bayésienne développée dans cette thèse pour des modèles où les observations ne sont pas indépendantes (par exemple, faiblement dépendantes).*

5. L'étude des modèles hétéroscedastiques (i.e. la variance n'est pas constante, $Y_i = f(X_i) + \sigma(X_i)\xi_i$) est une perspective intéressante. Nous pensons que l'estimateur bayésien peut être appliqué à ce genre de modèles.

   Problème Ouvert 5. *Peut-on utiliser l'estimateur bayésien dans des modèles hétéroscedastiques ?*

6. Dans notre travail, nous nous sommes restreints aux espaces isotropes. Cela est dû à l'adaptation et à la méthode de Lepski. Il est possible d'atteindre la vitesse $n^{-\frac{\bar{\beta}}{\gamma\bar{\beta}+1}}$ ($\bar{\beta}$ est la moyenne harmonique) sur les espaces de Hölder anisotropes avec l'estimateur bayésien dans le modèle de régression générale. En revanche, pour l'adaptation sur ces espaces, la méthode développée par Kerkyacharian, Lepski, et Picard [2001] ne convient pas pour les estimateurs non-linéaires.

   Problème Ouvert 6. *Existe-t-il une méthode adaptative capable d'estimer de façon optimale (atteindre la vitesse $\phi_{n,\gamma}(\bar{\beta})$) les fonctions anisotropes avec l'estimateur bayésien ?*

7. Le point de vue minimax adaptatif est souvent vu comme pessimiste. La recherche d'inégalités de type oracle permet d'éliminer ce problème.

   Problème Ouvert 7. *Obtenir des inégalités oracle pour la famille des estimateurs bayésiens indéxés par la fenêtre. Par exemple, l'approche de Goldenshluger et Lepski [2009a] peut convenir pour l'étude des fonctions isotropes.*

8. Les hypothèses 3 sont suffisantes pour montrer que l'estimateur bayésien est adaptatif et optimal (dans certains modèles). La question peut se formuler comme suit.

Problème Ouvert 8. *Existe-t-il des hypothèses nécessaires pour montrer que l'estimateur bayésien est adaptatif optimal ? Plus généralement, peut-on trouver les hypothèses nécessaires pour l'existence d'un estimateur adaptatif optimal ?*

9. Nous avons montré que l'estimateur bayésien atteignait la vitesse $\phi_{n,\gamma}$ pour la régression générale sous les hypothèses 3. De façon générale, il n'est pas prouvé que cette vitesse est optimale. On peut seulement montrer son optimalité modèle par modèle. Une perspective est de le démontrer pour n'importe quel modèle.

   Problème Ouvert 9. *Démontrer, avec le critère de Klutchnikoff [2005], que la vitesse $\phi_{n,\gamma}$ est optimale pour tout modèle sous les hypothèses 3.*

10. Un autre problème ouvert est celui des hypothèses. Nous constatons que les conditions sur la densité $g(.,.)$ se font à travers les hypothèses 3 sur le rapport de vraisemblance $Z_{n,\theta}$. Nous aimerions trouver les conditions nécessaires sur la densité (à l'aide de la distance d'Hellinger) pour construire un estimateur bayésien adaptatif optimal.

    Problème Ouvert 10. *Peut-on donner des hypothèses sur la densité g qui impliquent les hypothèses 3 ?*

11. L'approche bayésienne consiste à mettre une loi *a priori* sur le paramètre à estimer. Dans ce cas, le critère bayésien devient $\int \|t-u\| d_a(u)du$, où $d_a(u)$ remplace la vraisemblance $\big(L_h(u)\big)^{1/m}$ utilisée dans cette thèse. Il est possible de démontrer nos résultats pour n'importe quel *a priori* $d_a(u)$ sous les hypothèses 3.

    Problème Ouvert 11. *Remplacer $Z_{n,\theta}$ par un a priori et donner une méthode pour choisir l'a priori optimal sous les hypothèses 3 (voir par exemple Ghosal, Lember, et Van der Vaart [2008] et Van der Vaart et Van Zanten [2009]).*

2.4.2 Critère de Huber

L'approche de Huber est nouvelle pour l'adapatation. Seuls les résultats de Reiss, Rozenholc, et Cuenod [2009] traîtent du sujet pour l'estimateur localement constant et non localement polynômial avec un degré quelconque.

1. Pour cette approche, nous nous sommes aussi restreints aux espaces isotropes. En effet, les approches développées par Kerkyacharian, Lepski, et Picard [2001], Goldenshluger et Lepski [2008] et Goldenshluger et Lepski [2009a] ne conviennent pas pour les estimateurs non-linéaires (on ne peut pas comparer les biais avec des estimateurs non-linéaires). Nous avons espoir de trouver de nouvelles méthodes qui permettent de traiter les fonctions anisotropes. La recherche d'inégalités oracle est aussi une perspective.

Problème Ouvert 12. *Existe-t-il une méthode adaptative pour obtenir des inégalités oracle pour les fonctions anisotropes avec la famille des estimateurs de Huber indéxés par la fenêtre ?*

2. Comme pour l'approche bayésienne, nous pensons pouvoir utiliser l'approche de *Huber* pour les modèles hétéroscedastiques. En effet, le contrôle des grandes déviations repose sur l'inégalité de Bernstein qui peut être établie dans le cas hétéroscedastique. En revanche, l'indépendance des observations est absolument nécessaire pour utiliser cet estimateur.

   Problème Ouvert 13. *Peut-on utiliser l'estimateur de Huber dans des modèles hétéroscedastiques ?*

3. Nous pensons aussi que les hypothèses 1 peuvent être affaiblies. Notamment, il serait intéressant d'étudier le problème ouvert suivant.

   Problème Ouvert 14. *Peut-on obtenir les mêmes résultats avec l'approche de Huber avec seulement l'hypothèse de symétrie du bruit ? Dans le cas où le bruit est symétrique et $g(0) = 0$, peut-on trouver une vitesse de convergence meilleure que celle obtenue jusqu'ici par l'estimateur de Huber ? (l'information de Fisher est égale à l'infinie).*

4. Il est important de prouver que la vitesse $\phi_{n,2}$ (atteinte par l'estimateur de Huber), est optimale. Par exemple, pour le bruit gaussien ou de Cauchy, il est possible de le démontrer (avec les travaux de Klutchnikoff [2005], dans le cas minimax voir Tsybakov [2008], Chapitre 2). En revanche, dans le modèle du bruit additif uniforme, cette vitesse n'est pas optimale, car déjà dans le cas minimax, on a $n^{-\frac{\beta}{\beta+d}}$.

   Problème Ouvert 15. *A l'aide des travaux de Klutchnikoff [2005], donner les conditions nécessaires sur g_ξ, pour que la vitesse adaptative $\phi_{n,2}$ soit optimale.*

5. Remarquons que la fonction de Huber mélange la norme ℓ_1 et la norme ℓ_2. La norme ℓ_1 est là pour contrôler les valeurs extrêmes (cela utilise les avantages de la médiane) et la norme ℓ_2 est utilisée autour de 0 pour rendre le critère continu dans son comportement. La taille du voisinage autour de 0 où est utilisée la norme ℓ_2 peut être changée (dans notre cas, la taille est égale à 1). Cela ne change rien en théorie mais en pratique, ce voisinage est à calibrer. Les estimateurs robustes sont de plus en plus demandés en imagerie pour leur stabilité (voir Arias-Castro et Donoho [2009] ou Astola, Egiazarian, Foi, et Katkovnik [2010]). Développer cette approche pour l'imagerie (ou d'autres domaines) est une perspective intéressante.

Problème Ouvert 16. *En pratique, est-ce que la taille du voisinage, de la norme ℓ_2 dans le critère de Huber, améliore la performance ?*

6. Au vue des preuves du chapitre 5, il est possible d'utiliser d'autre critère que celui de Huber.

Problème Ouvert 17. *Trouver les conditions nécessaires sur le critère (par exemple, convexité, dérivée bornée, etc.) pour construire un estimateur adaptatif qui atteigne la vitesse $\phi_{n,2}$ dans le modèle de régression additive (voir Section 1.2.3).*

Chapter 3

General Locally Bayesian Approach

Ce chapitre traite du modèle de régression générale présenté dans la section 1.2. Nous développons un *estimateur Bayésien* déjà présenté en section 1.3.1. Le contenu de ce chapitre peut être trouvé dans l'article de Chichignoud [2010b], qui généralise le chapitre 4 et l'article de Chichignoud [2010a]. Sous certaines hypothèses 1 sur la densité du modèle, nous montrerons que notre *estimateur Bayésien* est optimal au sens minimax et minimax adaptatif (voir sections 3.2 et 3.3). Dans la section 3.4 nous présentons des exemples de modèles de régression, nous donnons en particulier les preuves qui permettent de vérifier les hypothèses 1 pour ces exemples et ainsi établir les vitesses de convergences minimax pour chaque modèle. Le contenu de ce chapitre généralise les résultats du chapitre 4. Les résultats de ce chapitre permettent de conclure le fait suivant : la famille des *estimateurs Bayésiens* est plus "riche" que les *estimateurs linéaires*.

3.1 Introduction

Let statistical experiment

$$\mathfrak{G}_n = \left\{ \left[[0,1]^d \times \mathbb{R} \right]^n, \ \mathbb{B}^{(n)}, \mathbb{P}_f^{(n)}, \ f \in \mathfrak{F} \right\}, \quad n \in \mathbb{N}^*,$$

be generated by independent random observations

$$\mathcal{Z}_n = (X_i, Y_i)_{i=1,\dots n}, \tag{3.1.1}$$

where the d-dimensional vector $X_i \in [0,1]^d$ can be viewed as a location in time or space and Y_i as the "observation at X_i".Here $f : [0,1]^d \to \mathbb{R}$ is unknown function and we are interested in estimating f at a given point $y \in [0,1]^d$ from observation $\mathcal{Z}_n$.

The design points $(X_i)_{i\in 1,\dots,n}$ are deterministic and without loss of generality, we will assume that

$$X_i \in \left\{ 1/n^{1/d}, 2/n^{1/d}, \dots, 1 \right\}^d, \ \ i = 1, \dots, n.$$

The design can be considered random and the proof is the same. But, we choose the deterministic design to simplify the notation.

Our model assumes that the values X_i are given and a distribution $g(., f_i)$ of each Y_i is determined by a parameter f_i, which may depend on the location X_i, $f_i = f(X_i)$. The estimation problem is to reconstruct $f(y)$ from the data $\mathcal{Z}_n$. Let us illustrate this set-up by few special cases.

1. *Gaussian regression.* Let $Z_i = (X_i, Y_i)$ with $X_i \in [0,1]^d$ and $Y_i \in \mathbb{R}$ obeying the regression equation $Y_i = f(X_i) + \xi_i$ with a regression function f and i.i.d. Gaussian errors $\xi_i \sim \mathcal{N}(0,1)$.
2. *Inhomogeneous Poisson regression.* Here the observation is discrete $Y_i \in \mathbb{N}$ and $X_i \in [0,1]^d$ is deterministic. The law of probability of Y_i can be written
$$g(k, f_i) = \mathbb{P}_f(Y_i = k) = \frac{[f(X_i)]^k}{k!} \exp\left\{-f(X_i)\right\}, \quad k \in \mathbb{N}.$$
This is the Poisson distribution with parameter $f(X_i)$. The goal is to estimate the function f at a given point y.
3. α *Regression* Let $0 < \alpha < 1/2$ and $Z_i = (X_i, Y_i)$ satisfied the regression equation $Y_i = f(X_i) + \epsilon_i$ where $(\epsilon_i)_i$ are i.i.d. random variables of density $g(x) = C(\alpha)\, e^{-|x|^\alpha}$, $x \in \mathbb{R}$.
4. *Multiplicative uniform regression.* Here $Z_i = (X_i, Y_i)$ satisfied the regression equation $Y_i = f(X_i) \times U_i$ where $U_i \sim \mathcal{U}_{\left[0,1\right]}$.

Along this thesis, the unknown function f is supposed to be smooth, in particular, it belongs to the Hölder ball of isotropic functions $\mathfrak{F} = \mathbb{H}_d(\beta, L, M)$ (see Definition 11 below). Here $\beta > 0$ is the smoothness of f, M is the upper bound of f and its partial derivatives and $L > 0$ is Lipshitz constant. We assume that we know *the minimax rate of convergence* (See Definition 12) given by the sequence

$$\varphi_{n,\gamma}(\beta) = n^{-\frac{\beta}{\gamma\beta+d}}, \tag{3.1.2}$$

where the parameter γ is known and depend on the function $g(.)$. For example, if $\mathcal{Z}_n$ satisfied the *gaussian regression* then $\gamma = 2$ and if $\mathcal{Z}_n$ satisfied the *multiplicative uniform regression* then $\gamma = 1$.

Minimax estimation. The first part of this chapter is devoted to the minimax over $\mathbb{H}_d(\beta, L, M)$ estimation. This means, in particular, that the parameters β, L and M are supposed to be known *a priori*. We propose the estimator being optimal in minimax sense (see Definition 12), i.e. the estimator attaining the normalising sequence (3.1.2). To construct the minimax estimator we use so-called *locally bayesian estimation construction* which consists in the following. Let

$$V_h(y) = \left\{ i = 1, ..., n : \; X_i \in [0,1]^d \cap \bigotimes_{j=1}^{d} \left[y_j - h/2, y_j + h/2\right] \right\},$$

be the neighborhood around y, where $h \in (0,1)$ be a given scalar. Fix $b > 0$ (without loss of generality we will assume that b is integer) and let

$$D_b = \sum_{m=0}^{b} \binom{m+d-1}{d-1}.$$

Let $K(z), z \in \mathbb{R}^d$ be the D_b-dimensional vector of polynomials of the following type (the sign $\top$ below means the transposition):

$$K^\top(z) = \left(\prod_{j=1}^{d} z_j^{p_j}, \ \ (p_1, \dots, p_d) \in \mathbb{N}^d : \ 0 \le p_1 + \dots + p_d \le b \right).$$

For any $t \in \mathbb{R}^{D_b}$ we define the local polynomial

$$f_t(x) \ = \ t^\top K\left(\frac{x-y}{h}\right) \mathbb{I}_{V_h(y)}(x), \quad x \in [0,1]^d, \tag{3.1.3}$$

where $\mathbb{I}$ denotes the indicator function. The local polynomial f_t can be viewed as an approximation of the regression function f inside of the neighborhood $V_y(h)$. Introduce the following subset of $\mathbb{R}^{D_b}$

$$\Theta\big(M\big) = \big\{ t \in \mathbb{R}^{D_b} : \ \ \|t\|_1 \le M \big\}, \tag{3.1.4}$$

where $\|.\|_1$ is ℓ_1-norm on $\mathbb{R}^{D_b}$. For any $i = 1, ..., n$ remark that $g\big(., f(X_i)\big)$ is the density of the observation Y_i. Let $\mathbb{E}_f = \mathbb{E}_f^n$ be the mathematical expectation with respect to the probability law $\mathbb{P}_f = \mathbb{P}_f^n$ of the observation $\mathcal{Z}_n$. Consider the *pseudo likelihood ratio*

$$L_h\left(t, \mathcal{Z}_n\right) = \prod_{X_i \in V_h(y)} g\big(Y_i, f_t(X_i)\big), \quad t \in \Theta\big(M\big), \tag{3.1.5}$$

Set also

$$\pi_h(t) = \int_{\Theta(M)} \|t-u\|_1 \ \left[L_h\big(u, \mathcal{Z}_n\big)\right]^{1/m} du, \quad t \in \Theta\big(M\big). \tag{3.1.6}$$

where m is a positive constant chosen in Assumptions 1. Let $\hat{\theta}(h)$ be the solution of the following minimization problem:

$$\hat{\theta}(h) = \arg \min_{t \in \Theta(M)} \pi_h(t). \tag{3.1.7}$$

The *locally bayesian estimator* $\bar{f}^h(y)$ of $f(y)$ is defined now as $\bar{f}^h(y) = \hat{\theta}_{0,\dots,0}(h)$.

We note that similar locally parametric approach based on maximum likelihood estimators was recently proposed in Polzehl and Spokoiny [2006] and Katkovnik and Spokoiny [2008] for *regular statistical models*.

As we see our construction contains an extra-parameter h to be chosen. To make this choice, we use quite standard arguments. First, we note that in view of $f \in \mathbb{H}_d(\beta, L, M)$

$$\exists \theta = \theta(f, y, h) \in \Theta\big(M\big) : \quad \sup_{x \in V_h(y)} \big|f(x) - f_\theta(x)\big| \le Ldh^\beta.$$

Thus, if h is chosen sufficiently small our original observation (3.1.1) is well approximated inside of $V_h(y)$ by the "parametric" observation $\mathcal{Y}_i$ with the density $g\big(. - f_\theta(X_i)\big)$ in which the *bayesian estimator* $\hat{\theta}$ is rate-optimal (See Theorem 4).

Finally, $h_n(\beta, L)$ is chosen as the solution of the following minimization problem

$$N_h^{-1} + Ldh^\beta \to \min_h \tag{3.1.8}$$

and we show that corresponding estimator $\bar{f}^{h_n(\beta,L)}(y)$ is minimax for $f(y)$ on $\mathbb{H}_d(\beta, L, M)$ if $\beta \le b$. Since the parameter $b > 0$ can be chosen on arbitrary way, the proposed estimator is minimax for any given value of the parameter $\beta > 0$.

Adaptive estimation. The second part of this chapter is devoted to the adaptive minimax estimation over collection of isotropic functional classes. To our knowledge, the problem of optimal adaptive estimation in the general regression, with no-linear estimators, is not studied in the literature.

Well-known drawback of minimax approach is the dependence of the minimax estimator on the parameters describing functional class on which the maximal risk is determined. In particular, the locally bayesian estimator $\bar{f}^h(\cdot)$ depends obviously on the parameter M via the solution of the minimization problem (3.1.7). Moreover, $h_n(\beta, L)$ optimally chosen in view of (3.1.8) depends explicitly on β and L. To overcome this drawback the minimax adaptive approach was proposed (see Lepski [1990], Lepski [1991], Lepski, Mammen, and Spokoiny [1997]). The first question arising in the adaptation (reduced to the problem at hand) can be formulated as follows.

Does there exist an estimator which would be minimax on $\mathbb{H}_d(\beta, L, M)$ simultaneously for all values of β, L and M belonging to some given subset of $\mathbb{R}_+^3$?

In section 3.3, we determine that the answer on this question is negative, that is typical for the estimation of the function at a given point (Lepski and Spokoiny [1997]). This answer can be reformulated in the following manner: the family of rates of convergence $\big\{\varphi_{n,\gamma}(\beta),\ \beta \in \mathbb{R}_+^*\big\}$ is unattainable for the problem under consideration.

Thus, we need to find another family of normalizations for maximal risk which would be attainable and, moreover, optimal in view of some criterion of optimality. Nowadays, the most developed criterion of optimality is due to Klutchnikoff [2005].

We can show that the family of normalizations, being optimal in several models (gaussian

$\gamma = 2$, uniform $\gamma = 1$ or α regression $\gamma = 1 + 2\alpha$) in view of this criterion, is

$$\phi_{n,\gamma}(\beta) = \left(\frac{\rho_{n,\gamma}(\beta)}{n}\right)^{\frac{\beta}{\gamma\beta+d}}, \qquad \rho_{n,\gamma}(\beta) = \left[1 + \frac{\gamma(b-\beta)}{(\gamma b + d)(\gamma\beta + d)} \ln n\right]^{\frac{1}{\gamma}}, \tag{3.1.9}$$

whenever $\beta \in]0, b]$. The factor $\rho_{n,\gamma}$ can be considered as *price to pay for adaptation* (Lepski [1990]).

The most important step in proving the optimality of the family (3.1.9) is to find an estimator, called *adaptive*, which attains the optimal family of normalizations. Obviously, we seek an estimator whose construction is *parameter-free*, i.e. independent of β and L. In order to explain our estimation procedure, let us make several remarks.

First, we note that the role of the constants β, L in the construction of the minimax estimator is quite different. Indeed, the constant M is used in order to determine the set $\Theta\big(M\big)$ needed for the construction of the locally bayesian estimator, see (3.1.6) and (3.1.7). However, this set does not dependent on the localisation parameter $h > 0$, in other words, the quantity M is not involved in the selection of optimal size of the local neighborhood given by (3.1.8). Contrary to that, the constants β, L are used for the derivation of the optimal size of the local neighborhood (3.1.8), but they are not involved in the construction of the collection of locally bayesian estimators $\big\{\hat{f}^h, h > 0\big\}$.

In order to select an "optimal" estimator from the family $\big\{\hat{f}^h, h > 0\big\}$ we use general adaptation scheme due to Lepski [1990,1992a]. Lepski's procedure is typically applied to the selection from the collection of linear estimator (kernel estimators, locally polynomial estimator, etc.). In the present paper, we apply this method to a very complicated family of non-linear estimators, obtained by the use of bayesian approach on the random parameter set. It required, in particular, to establish the exponential inequality for the deviation of locally bayesian estimator from the parameter to be estimated (Proposition 3). It generalizes the inequality proved for the parametric model (Has'minskii and Ibragimov [1981]). This results seems to be new.

3.2 Minimax Estimation

In this section we present a upper bound concerning minimax estimation. First, we establish lower bound for minimax risk defined on $\mathbb{H}_d(\beta, L, M)$ for any β, L and M. For any $(p_1, ..., p_d) \in \mathbb{N}^d$ we denote $\vec{p} = (p_1, ..., p_d)$ and $|\vec{p}| = p_1 + ... + p_d$.

Definition 11. *Fix $\beta > 0$, $L > 0$ and $M > 0$ and let $\lfloor\beta\rfloor$ be the largest integer strictly less than β. The* isotropic Hölder class $\mathbb{H}_d(\beta, L, M)$ *is the set of functions $f : [0,1]^d \to \mathbb{R}$ having on $[0,1]^d$ all partial derivatives of order $\lfloor\beta\rfloor$ and such that $\forall x, y \in [0,1]^d$*

$$\sum_{m=0}^{\lfloor\beta\rfloor}\sum_{|\vec{p}|=m}\sup_{x\in[0,1]^d}\left|\frac{\partial^{|\vec{p}|}f(x)}{\partial x_1^{p_1}\cdots\partial x_d^{p_d}}\right| \leq M,$$

$$\left|\frac{\partial^{|\vec{p}|}f(x)}{\partial x_1^{p_1}\cdots\partial x_d^{p_d}}-\frac{\partial^{|\vec{p}|}f(y)}{\partial y_1^{p_1}\cdots\partial y_d^{p_d}}\right| \leq L\big[\|x-y\|_1\big]^{\beta-\lfloor\beta\rfloor},\quad \forall\,|\vec{p}|=\lfloor\beta\rfloor.$$

Maximal and Minimax Risk on $\mathbb{H}_d(\beta,L,M)$. To measure the performance of estimation procedures on $\mathbb{H}_d(\beta,L,M)$ we will use minimax approach.

First, we define the maximal risk on $\mathbb{H}_d(\beta,L,M)$ corresponding to the estimation of the function f at a given point $y\in[0,1]^d$.
Let $\tilde{f}$ be an arbitrary estimator built from the observation $\mathcal{Z}_n$. Let $\forall q>0$

$$R_{n,q}\big[\tilde{f},\mathbb{H}_d(\beta,L,M)\big]=\sup_{f\in\mathbb{H}_d(\beta,L,M)}\mathbb{E}_f\big|\tilde{f}(y)-f(y)\big|^q.$$

The quantity $R_{n,q}\big[\tilde{f},\mathbb{H}_d(\beta,L,M)\big]$ is called *maximal risk* of the estimator $\tilde{f}$ on $\mathbb{H}_d(\beta,L,M)$ and the *minimax risk* on $\mathbb{H}_d(\beta,L,M)$ is defined as

$$R_{n,q}\big[\mathbb{H}_d(\beta,L,M)\big]=\inf_{\tilde{f}}R_{n,q}\big[\tilde{f},\mathbb{H}_d(\beta,L,M)\big],$$

where inf is taken over the set of all estimators.

Definition 12. *The normalising sequence ψ_n is called minimax rate of convergence and the estimator $\hat{f}$ is called minimax (asymptotically minimax) if*

$$\liminf_{n\to\infty}\psi_n^{-q}\,R_{n,q}\big[\hat{f},\mathbb{H}_d(\beta,L,M,A)\big] > 0;$$
$$\limsup_{n\to\infty}\psi_n^{-q}\,R_{n,q}\big[\hat{f},\mathbb{H}_d(\beta,L,M,A)\big] < \infty.$$

We remember that the first assumption of this chapter is that $\varphi_{n,\gamma}$ defined in (3.1.2) is the *minimax rate of convergence*. To obtain the upper bound of minimax risk, we need to the following assumptions.

Let us introduce the following notations. For any $\theta\in\Theta(M)$, let $U_n=N_h\big[\Theta\big(M\big)-\theta\big]$. Put also $N_h=N_h(\gamma)=\big(nh^d\big)^{1/\gamma}$ where $\gamma\geq 1$ and $\forall u\in U_n$

$$Z_{n,\theta}(u)=\frac{L_h\big(\theta+u\,N_h^{-1},\mathcal{Z}_n\big)}{L_h\big(\theta,\mathcal{Z}_n\big)}. \tag{3.2.1}$$

Let $\mathcal{H}_n, n>1$ be the following subinterval of $(0,1)$.

$$\mathcal{H}_n=[h_{\min},\,h_{\max}],\quad h_{\min}=\big(\ln n\big)^{\frac{1}{\gamma d+d^2}}n^{-1/d},\quad h_{\max}=(\ln n)^{-\frac{1}{\gamma b+d}}. \tag{3.2.2}$$

Later on we will consider only the values of h belonging to $\mathcal{H}_n$. Let $f_\theta(x)$, with $\theta \in \Theta\big(M\big)$, be the local polynomial approximation of f inside $V_h(y)$ and let b_h be the corresponding approximation error, i.e.

$$b_h = \sup_{x \in V_h(y)} \big|f_\theta(x) - f(x)\big|. \tag{3.2.3}$$

Put finally $\mathcal{N}(h) = \big(b_h \times N_h\big)^\gamma$ and define

$$\mathcal{E}_h = \exp\left\{\mathcal{N}(h)\right\}. \tag{3.2.4}$$

Let the subset $\Gamma_\delta \subseteq \Theta\big(M\big)$ such that $\int_{\Gamma_\delta} du = \delta^{D_b}$, $\delta > 0$.

Assumptions 1. *We suppose that it exists several constants $\tau, m > 0$ and c_1, c_2, c_3, C_1, C_2, s_1, $s_2 > 0$ such that for any $\forall n > 1$, $f \in \mathbb{H}_d(\beta, L, M)$, $\theta \in \Theta\big(M\big)$ and $h \in \mathcal{H}_n$, we have*

1. $\displaystyle\int_{\Gamma_\delta} \mathbb{E}_f \left[1 - Z_{n,\theta}^{1/m}(u)\right]_+ du \le C_1\, \mathcal{E}_h^{c_1} \delta^{D_b+\tau}, \quad \forall \delta < s_1,$
2. $\mathbb{E}_f Z_{n,\theta}^{1/m}(u) \le C_2\, \mathcal{E}_h^{c_2}\ \exp\left\{-c_3\, \|u\|_1^\gamma\right\}, \ \ \forall u \in U_n : \ \|u\|_1 \ge s_2\, \mathcal{N}(h).$

where $a_+ = \max(a, 0)$, $a \in \mathbb{R}$. We also assume the acknowledgement of constants c_3, s_2, τ, m and γ.

The next theorem shows how to construct the minimax estimator basing on locally bayesian approach. Put $\bar{h} = \big(L^\gamma n\big)^{-\frac{1}{\gamma\beta+d}}$ and let $\bar{f}^{\bar{h}}(y) = \hat{\theta}_{0,\dots,0}\big(\bar{h}\big)$ is given by (3.1.4), (3.1.6) and (3.1.7) with $h = \bar{h}$.

Theorem 4. *Let $\beta > 0$, $L > 0$ and $M > 0$ be fixed. Assume that Assumptions 1 are verified. Then it exists the constant C_* such that for any $n \in \mathbb{N}^*$ satisfying $N_{\bar{h}} \ge 1$,*

$$\varphi_{n,\gamma}^{-q}(\beta) R_{n,q}\Big[\bar{f}^{\bar{h}}(y), \mathbb{H}_d(\beta, L, M)\Big] \le C_*, \quad \forall q \ge 1.$$

3.3 Adaptive Rule

This section is devoted to the adaptive estimation over the collection of the classes $\Big\{\mathbb{H}_d(\beta, L, M)\Big\}_{\beta,L,M}$. We will not impose any restriction on possible values of L, M, but we will assume that $\beta \in (0, b]$, where b, as previously, is an arbitrary *a priori* chosen integer.

Let Φ be the following family of normalizations:

$$\phi_{n,\gamma}(\beta) = \left(\frac{\rho_{n,\gamma}(\beta)}{n}\right)^{\frac{\beta}{\gamma\beta+d}}, \quad \rho_{n,\gamma}(\beta) = \left[1 + \frac{\gamma(b-\beta)}{(\gamma b + d)(\gamma\beta + d)} \ln n\right]^{\frac{1}{\gamma}}, \ \ \beta \in (0, b].$$

We remark that $\phi_{n,\gamma}(b) = \varphi_{n,\gamma}(b)$ and $\rho_{n,\gamma}(\beta) \sim \big(\ln n\big)^{\frac{1}{\gamma}}$ for any $\beta \ne b$.

Construction of Φ-Adaptive Estimator. As it was already mentioned in Introduction the construction of our estimation procedure consists of two steps. First, we define the family of locally bayesian estimators. Next, based on Lepski's method, we propose data-driven selection from this family.

First step: Collection of locally bayesian estimators. Put

$$h_k = 2^{-k} h_{\max}, \quad k = 0, \dots, \mathbf{k}_n,$$

where $\mathbf{k}_n$ is largest integer such that $h_{\mathbf{k}_n} \in \mathcal{H}_n$ defined in (3.2.2). Set

$$\hat{\mathcal{F}} = \left\{ \hat{f}^{(k)}(y) = \hat{\theta}_{0,\dots,0}(h_k), \quad k = 0, \dots, \mathbf{k}_n \right\},$$

where $\hat{\theta}_{0,\dots,0}(h_k)$ is given by (3.1.4), (3.1.6) and (3.1.7) with $h = h_k$.

Second step: Data-driven selection from the collection $\hat{\mathcal{F}}$. We put $\hat{f}^*(y) = \hat{f}^{(\hat{k})}(y)$, where $\hat{f}^{(\hat{k})}(y)$ is selected from $\hat{\mathcal{F}}$ in accordance with the rule:

$$\hat{k} = \inf \left\{ k = \overline{0, \mathbf{k}_n} : \ \left| \hat{f}^{(k)}(y) - \hat{f}^{(l)}(y) \right| \le C S_n(l), \ \ l = \overline{k+1, \mathbf{k}_n} \right\}. \tag{3.3.1}$$

Here we have used the following notation.

$$C = \left(s_2^\gamma \vee \frac{2^{2\gamma+1}(2\gamma + 2dq)}{c_3 \gamma \left(1 \wedge \tau D_b^{-1}\right)} \right)^{\frac{1}{\gamma}}, \quad S_n(l) = \left[\frac{1 + l \ln 2}{n (h_l)^d} \right]^{\frac{1}{\gamma}}, \quad l = 0, 1, \dots, k_n. \tag{3.3.2}$$

Theorem 5. *Let $b > 0$ be fixed. Assume that Assumptions 1 are verified, then for any $\beta \in (0, b]$, $L > 0$, $M > 0$ and $q \ge 1$*

$$\limsup_{n \to \infty} \phi_{n,\gamma}^{-q}(\beta) R_{n,q}\left[\hat{f}^*(y), \mathbb{H}_d(\beta, L, M) \right] < \infty.$$

Remark 7. *The assertion of the theorem means that the proposed estimator $\hat{f}^*(y)$ is Φ-adaptive. It implies in particular that the family of normalizations Φ is admissible. For example (Section 4.6), we can state the optimality of Φ in view of Klutchnikoff criterion.*

3.4 Applications

In this section, we show how the upper bounds of Theorems 4 and 5 can be used for the derivation of minimax and adaptive minimax results for different additive models of type 3.1.1. In particular, in following sections, we consider four particular problems:

1. $g(. - f_i)$ is the normal distribution $\mathcal{N}(0, 1)$,

2. $g(k, f_i) = \mathbb{P}_f(Y_i = k) = \frac{[f(X_i)]^k}{k!} \exp\{-f(X_i)\}, \quad k \in \mathbb{N},$
3. $g(x, f_i) = C(\alpha) \exp\left\{-\left|x - f_i\right|^\alpha\right\}, \quad x \in \mathbb{R},$
4. $g(., f_i)$ is the uniform distribution on $[0, f_i]$.

Our goal here is to show how the assumption on $g(.)$ leads to bayesian estimators with optimal statistical properties. Note that in each particular case for adaptation, the bayesian estimators are obtained by the same selection rule presented in Section 3.3. Here, we always assume that the unknown $f \in \mathbb{H}_d(\beta, L, M)$.

3.4.1 Gaussian Regression

Consider the model 3.1.1 with $g(x, f_i) = \left(\sqrt{2\pi}\right)^{-1} \exp\{-(x - f_i)^2/2\}$, we obtain the following model:

$$Y_i = f(X_i) + \xi_i, \quad i = 1, ..., n. \tag{3.4.1}$$

According to Tsybakov [2008], it is well-known that the minimax rate is $n^{-\frac{\beta}{2\beta+d}}$, thus here we take $\gamma = 2$. By definition (3.2.1) the process $Z_{n,\theta}$ can be formulate like

$$Z_{n,\theta}(u) = \prod_{X_i \in V_y(h)} \exp\left\{-2^{-1}\left(Y_i - f_{\theta + uN_h^{-1}}(X_i)\right)^2 + 2^{-1}\left(Y_i - f_\theta(X_i)\right)^2\right\}$$

Put $\lambda_n(h)$ the smallest eigenvalue of the matrix

$$\mathcal{M}_{nh}(y) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{X_i - y}{h}\right) K^\top \left(\frac{X_i - y}{h}\right) \mathbb{I}_{V_h(y)}(X_i), \tag{3.4.2}$$

which is completely determined by the design points and by the number of observations. We will prove that it exists $\lambda > 0$, such that $\lambda_n(h) \geq \lambda$ for any $n \geq 1$ and any $h \in \left[h_{\min}, h_{\max}\right]$ (see Lemma 9 in Chapter 4). Let us give the following lemma which verified Assumptions 1 on $Z_{n,\theta}$.

Lemma 1. *For any $f \in \mathbb{H}_d(\beta, L, M)$, $n \geq 1$, $h \in \mathcal{H}_n$, $\theta \in \Theta(M)$*

1. $\displaystyle\int_{[0,\delta]^{D_b}} \mathbb{E}_f \left[1 - Z_{n,\theta}^{1/2}(u)\right]_+ du \leq 2\mathcal{E}_h \delta^{D_b+2}, \quad \forall \delta < 1,$
2. $\displaystyle\mathbb{E}_f Z_{n,\theta}^{1/2}(u) \leq \exp\left\{-\frac{D_b \lambda_n(h)}{16} \|u\|_1^2\right\}, \quad \forall u \in U_n : \ \|u\|_1 \geq \frac{8}{D_b \lambda_n(h)} \mathcal{N}(h).$

The proof is given in Appendix. We use the adaptive bayesian estimator noted $\hat{f}_1^*(y)$ and developed in Section 3.3 with model 3.4.1, by identification we take constants $\gamma = 2$, $\tau = 2$, $m = 2$, $s_2 = 8\, D_b^{-1} \lambda_n^{-1}(h)$, $c_3 = 2^{-4} D_b\, \lambda_n(h)$.

Theorem 6. *Let $b > 0$ be fixed, then for any $\beta \in (0, b]$, $L > 0$, $M > 0$ and $q \geq 1$*

$$\limsup_{n\to\infty} \phi_{n,2}^{-q}(\beta) R_{n,q}\Big[\hat{f}_1^*(y), \mathbb{H}_d(\beta, L, M)\Big] < \infty.$$

Remark 8. *The proof of the theorem is omitted; it is a straightforward consequence of Theorem 5 and Lemma 1. We can deduced the minimax result to take $b = \beta$ and remark that $\phi_{n,2}(b) = \varphi_{n,2}(b)$. With Klutchnikoff criterion, we can prove that $\{\phi_{n,2}(\beta)\}_{\beta\in]0,b]}$ is adaptive optimal for the Gaussian regression and the pointwise estimation (See Klutchnikoff [2005]).*

3.4.2 Inhomogeous Poisson Regression

Consider the model 3.1.1 with $g(k, f_i) = \mathbb{P}_f(Y_i = k) = \frac{[f(X_i)]^k}{k!} \exp\{-f(X_i)\}, \quad k \in \mathbb{N}$. Assume that $f \in \mathbb{H}\big(\beta, L, M, A\big)$, $M \geq A > 0$, where

$$\mathbb{H}\big(\beta, L, M, A\big) = \Big\{ f \in \mathbb{H}\big(\beta, L, M\big) \ : \ \inf_{x\in[0,1]^d} f(x) \geq A \Big\}.$$

By definition (3.2.1) the process $Z_{n,\theta}$ can be formulate like

$$Z_{n,\theta}(u) = \prod_{X_i \in V_y(h)} \left(1 + \frac{f_{uN_h^{-1}}(X_i)}{f_\theta(X_i)}\right)^{Y_i} \exp\Big\{-f_{uN_h^{-1}}(X_i)\Big\}, \quad N_h = \big(nh^d\big)^{1/2}.$$

Put $\lambda_n(h)$ given in the previous section. Let us give the following lemma which verified Assumptions 1 on $Z_{n,\theta}$. Let us introduce the following notation

$$\Theta\big(A, M\big) = \big\{ t \in \mathbb{R}^{D_b} : \ \ 2t_{0,\dots,0} - \|t\|_1 \geq A, \ \ \|t\|_1 \leq M \big\}. \tag{3.4.3}$$

Lemma 2. *For any $f \in \mathbb{H}_d(\beta, L, M)$, $n \geq n_0$, $h \in \mathcal{H}_n$, $\theta \in \Theta\big(A, M\big)$*

1. $\displaystyle\int_{[0,\delta]^{D_b}} \mathbb{E}_f \Big[1 - Z_{n,\theta}^{1/2}(u)\Big]_+ du \leq \sqrt{\frac{2 + M + 2A}{2A}} \mathcal{E}_h^{1/2} \delta^{D_b + 1/2}, \quad \forall \delta < 1,$
2. $\mathbb{E}_f Z_{n,\theta}^{1/2}(u) \leq e^1 \exp\left\{-\dfrac{\lambda_n(h) A}{8M^2} \|u\|_1^2\right\}, \ \ \forall u \in U_n : \ \|u\|_1 \geq \dfrac{4M^2}{\lambda_n(h)} \mathcal{N}(h).$

Here, $U_n = N_h\big(\Theta\big(A, M\big) - \theta\big)$.

The proof is given in Appendix. Remark that this result is given for $n \geq n_0$, this do not change the final result which is asymptotic. We use the adaptive bayesian estimator noted $\hat{f}_2^*(y)$ and developed in Section 3.3, by identification we take constants $\gamma = 2$, $\tau = 0.5$, $m = 2$, $s_2 = 4M^2\lambda_n^{-1}(h)$, $c_3 = 2^{-3}M^{-2}\lambda_n(h)A$.

Theorem 7. *Let $b > 0$ be fixed, then for any $\beta \in (0, b]$, $L > 0$, $M \geq A > 0$ and $q \geq 1$*

$$\limsup_{n\to\infty} \phi_{n,2}^{-q}(\beta) R_{n,q}\left[\hat{f}_2^*(y), \mathbb{H}_d(\beta, L, M, A)\right] < \infty.$$

Remark 9. *The proof of the theorem is omitted; it is a straightforward consequence of Theorem 5 and Lemma 2. We can deduced the minimax result to take $b = \beta$ and remark that $\phi_{n,2}(b) = \varphi_{n,2}(b)$. A good exercice is to prove that $\{\phi_{n,2}(\beta)\}_{\beta\in]0,b]}$ is adaptive optimal for the Poisson regression and the pointwise estimation (See Klutchnikoff [2005]). In this case, the constant c_3 depends explicitly of A, M, unknown in practice. We can construct no-optimal estimators of A, M to choose the constant C in the Lepski's procedure. Few examples of estimators are given for the multiplicative uniform regression (Chapter 4).*

3.4.3 α Regression

We call α regression the following model

$$Y_i = f(X_i) + \epsilon_i, \quad i = 1, ..., n, \tag{3.4.4}$$

where ϵ_i has for density $g_\alpha(x) = C(\alpha) \exp\{-|x|^\alpha\}$ with $0 < \alpha < 1/2$. $C(\alpha)$ is a constant chosen so that g_α density is well. Here the design X_i is deterministic uniform on $[0, 1]^d$ and $f \in \mathbb{H}_d(\beta, L, M)$.

In this model, we want to achieve the rate of convergence $n^{-\frac{\beta}{(1+2\alpha)\beta+d}}$ (given by Has'minskii and Ibragimov [1981] in the parametric estimation). Thus, we take $\gamma = 1+2\alpha$. By definition (3.2.1) the process $Z_{n,\theta}$ can be formulate like

$$Z_{n,\theta}(u) = \prod_{X_i\in V_y(h)} \exp\left\{-\left|Y_i - f_{\theta+uN_h^{-1}}(X_i)\right|^\alpha + \left|Y_i - f_\theta(X_i)\right|^\alpha\right\}$$

Put $\lambda_n(h)$ given in Section 3.4.1. Let us give the following lemma which verified Assumptions 1 on $Z_{n,\theta}$.

Lemma 3. *For any $f \in \mathbb{H}_d(\beta, L, M)$, it exists c_1, $\mathcal{C}$, n_0 such that for $n \geq n_0$, $h \in \mathcal{H}_n$, $\theta \in \Theta(M)$, we have*

1. $\displaystyle\int_{[0,\delta]^{D_b}} \mathbb{E}_f\left[1 - Z_{n,\theta}^{1/2}(u)\right]_+ du \leq \mathcal{C}\mathcal{E}_h^{c_1}\delta^{D_b+\alpha+1/2}, \quad \forall \delta < 1,$

2. $\mathbb{E}_f Z_{n,\theta}^{1/2}(u) \leq \mathcal{E}_h^{\frac{1+C(\alpha)}{2(1+\alpha)}} \exp\left\{-c_3 \|u\|_1^{1+2\alpha}\right\}, \quad \forall u \in U_n,$

where $c_3 = \lambda_n(h) 2^{\alpha-4} C(\alpha)(1 + 2\alpha)^{-1} \exp\left\{-\frac{1}{2}(2M)^\alpha\right\}$.

The proof is given in Appendix. Remark that we do not know values of c_1, $\mathcal{C}$, it is not necessary to construct the adaptive estimator. We use the adaptive bayesian estimator noted $\hat{f}_3^*(y)$ and developed in Section 3.3 with model 3.4.4.

Theorem 8. *Let $b > 0$ be fixed, then for any $\beta \in (0, b]$, $L > 0$, $M > 0$ and $q \geq 1$*

$$\limsup_{n\to\infty} \phi_{n,1+2\alpha}^{-q}(\beta) R_{n,q}\Big[\hat{f}_3^*(y), \mathbb{H}_d(\beta, L, M)\Big] < \infty.$$

Remark 10. *The proof of the theorem is omitted, it is a straightforward consequence of Theorem 5 and Lemma 3. The optimality of $\phi_{n,1+2\alpha}(\beta)$ is not proved, it is a open problem. Here Lemma 3 is given with $n \geq n_0$, it will not be a problem because Theorem 8 is a asymptotic result. We can construct the minimax estimator with Theorem 4, because in the case minimax, the term $\mathcal{E}_h$ (in Assumptions 1) is a constant for $h = \bar{h}$.*

3.4.4 Multiplicative Uniform Regression

We consider the following model:

$$Y_i = f(X_i) \times U_i, \quad i = 1, ..., n, \tag{3.4.5}$$

where $(U_i)_i$ is a sequence independent identically distributed of uniform law on $[0, 1]$. This chapter 4 is devoted to the study of this model. So, we can find the minimax rate $n^{-\frac{\beta}{\beta+d}}$, thus, here we take $\gamma = 1$. The proof of Assumptions 1 is a consequence of Lemma 8 in Chapter 4 and we give the following lemma

Lemma 4. *For any $f \in \mathbb{H}_d(\beta, L, M, A)$, $n \geq 1$, $h \in \mathcal{H}_n$, $\theta \in \Theta\big(A, M\big)$*

1. $\displaystyle\int_{[0,\delta]^{D_b}} \mathbb{E}_f \Big[1 - Z_{n,\theta}^{1/2}(u)\Big]_+ du \leq \frac{3}{2A}\delta^{D_b+1}, \quad \forall \delta < 1,$

2. $\mathbb{E}_f Z_{n,\theta}^{1/2}(u) \leq \mathcal{E}_h^{1/A} \exp\left\{-\dfrac{\lambda_n(h)}{18MD_b}\|u\|_1\right\}, \quad \forall u \in N_h\big(\Theta\big(A, M\big) - \theta\big).$

where

$$Z_{n,\theta}(u) = \prod_{X_i \in V_y(h)} \frac{f_\theta(X_i)}{f_{\theta+uN_h^{-1}}(X_i)} \mathbb{I}_{\left\{Y_i \leq f_{\theta+uN_h^{-1}}(X_i)\right\}}.$$

This case requires us to several restrictions on the function f. Indeed, we assume that $f \geqq A > 0$ ($f \in \mathbb{H}_d(\beta, L, M, A)$) and the polynomial approximation is constructed with the constraint $f_\theta \geqq f$. We find a lot of details in Chapter 4, in particular this lemma is a consequence of Lemma 8.

We use the adaptive bayesian estimator noted $\hat{f}_4^*(y)$ and developed in Section 3.3 with model 3.4.5. By identification, we take constants $\gamma = 1$, $\tau = 1$, $m = 2$, $s_2 = 0$, $c_3 = \frac{\lambda_n(h)}{18MD_b}$.

Theorem 9. *Let $b > 0$ be fixed, then for any $\beta \in (0, b]$, $L > 0$, $M > 0$ and $q \geq 1$*

$$\limsup_{n\to\infty} \phi_{n,1}^{-q}(\beta) R_{n,q}\Big[\hat{f}_4^*(y), \mathbb{H}_d(\beta, L, M, A)\Big] < \infty.$$

Remark 11. *The proof of the theorem is omitted, it is a straightforward consequence of Theorem 5 and Lemma 4. With Klutchnikoff criterion, we can prove that* $\{\phi_{n,1}(\beta)\}_{\beta\in]0,b]}$ *is adaptive optimal for the multiplicative uniform regression and the pointwise estimation (See Theorem 13 in Chapter 4).*

3.5 Proofs of Main Results

In this section we give the proofs of Theorems 4 and 5.

3.5.1 Auxiliary Results: Large Deviations

For the sequel we fixe, for any $h > 0$ satisfying (3.2.2) and $f \in \mathbb{H}_d(\beta, L, M)$, $\theta = \theta(f, y, h) = \big\{\theta_{\vec{p}} : \vec{p} \in \mathbb{N}^d : 0 \le |\vec{p}| \le b\big\}$ such that $b_h \le Ldh^\beta$ and $\theta \in \Theta\big(M\big)$. For example, we can take θ the coefficients of expansion Taylor series.

The next proposition is the milestone for all results proved in the paper. Introduce the following notations $\omega_2 = c_1 2^{-1} \vee c_2$,

$$\omega_1 = 16\, C_2 \int_0^{+\infty} (z+1)e^{-c_3 z^\gamma} dz + 4\sqrt{C_1}, \quad \omega_3 = c_3 2^{-2\gamma-1}\left(1 \wedge \frac{\tau}{D_b}\right),$$

where $C_1,\ C_2,\ c_1,\ c_2,\ c_2,\ c_3$ are defined in section 3.2. Put

$$M(f) = \sum_{m=0}^{b} \sum_{p_1+\ldots+p_d=m} \left| \frac{\partial^m f(y)}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} \right|. \tag{3.5.1}$$

The following agreement will be used in the sequel: if the function f and $m \ge 1$ be such that $\partial^m f$ does not exist we will put formally $\partial^m f = 0$ in the definition of $M(f)$.

Proposition 3. *Let* $\varepsilon_0 = 4\, s_2 \vee \Big(c_3^{-1} 2^{2\gamma+1}\big(M \wedge \ln s_1\big)_+\Big)^{1/\gamma}$. *For any* $n > 1$, $h \in \mathcal{H}_n$ *and any* f *such that* $M(f) < \infty$, *then* $\forall \varepsilon \ge \varepsilon_0\, \mathcal{N}(h)$

$$\mathbb{P}_f\left(N_h \big|\hat{f}^h(y) - f(y)\big| \ge \varepsilon\right) \le \omega_1 \mathcal{E}_h^{\omega_2} \exp\left\{-\omega_3 \varepsilon^\gamma\right\},$$

where $\hat{f}^h(y)$ *is given by (3.1.4), (3.1.6) and (3.1.7) and* s_1, s_2 *are defined in Assumptions 1 in Section 3.2.*

The proof of this proposition is given in section 3.5.4.

3.5.2 Proof of Theorem 4

By integration of the inequality of Proposition 3, we obtain for any $n > 1$, $h \in \mathcal{H}_n$ and any f such that $M(f) < \infty$

$$\mathbb{E}_f|\hat{f}^h(y) - f(y)|^q \le C_q \mathcal{E}_h^{\left[q \vee (\omega_2 - \omega_3 \varepsilon_0/2)\right]} N_h^{-q}, \tag{3.5.2}$$

where $C_q = 2\varepsilon_0^q + \frac{4q\omega_1}{\omega_3} \sup_{\eta \ge 0} [\eta^{q-1} \exp\{-\omega_3 \eta^\gamma/2\}]$. By definition of $\bar{h} = (Ln)^{-\frac{1}{\beta+d}} \in \mathcal{H}_n$, we have

$$Ldnh^{\gamma\beta+d} = d, \quad N_{\bar{h}}^{-q} = L^{\frac{\gamma q d}{\gamma\beta+d}} \varphi_{n,\gamma}^q(\beta).$$

Using (3.5.2) and last equalities, we obtain for any $f \in \mathbb{H}_d(\beta, L, M)$

$$\mathbb{E}_f\big|\bar{f}^{\bar{h}}(y) - f(y)\big|^q \le C_q \, e^{d\left[q \vee (\omega_2 - \omega_3 \varepsilon_0/2)\right]} \, L^{\frac{qd\gamma}{\gamma\beta+d}} \, \varphi_{n,\gamma}^q(\beta).$$

Theorem 4 is proved. ∎

3.5.3 Proof of Theorem 5

We start the proof with formulating some auxiliary results whose proofs are given in Appendix 3.6. Define $J_1 = \omega_2$ and

$$h^* = \left[\frac{c\, \rho_{n,\gamma}^\gamma(\beta)}{L^\gamma d^\gamma n}\right]^{\frac{1}{\gamma\beta+d}}, \quad c < (2J_1)^{-\gamma}.$$

and let the integer κ be defined as follows.

$$2^{-\kappa} h_{\max} \le h^* < 2^{-\kappa+1} h_{\max}. \tag{3.5.3}$$

Lemma 5. *For any $f \in \mathbb{H}_d(\beta, L, M)$ and any $k \ge \kappa + 1$*

$$\mathbb{P}\big(\hat{k} = k\big) \le J_2 \exp\left\{J_1 n (h^*)^{\gamma\beta+d}\right\} 2^{-(k-1)\frac{2\gamma+2dq}{\gamma}},$$

where $J_2 = \omega_1 \left(1 - 2^{-\frac{2\gamma+2dq}{\gamma}}\right)^{-1}$.

Proof of Theorem 5. Note that $h^* \in \mathcal{H}_n$ and $h_l \in \mathcal{H}_n$, $l = 0, ..., \mathbf{k}_n$. The definition of h^* and κ in (3.5.3) implies that by integration of the inequality obtained in Proposition 3 with $h = h_k$, we obtain:

$$\mathbb{E}_f\big|\hat{f}^{(k)}(y) - f(y)\big|^q \le \bar{C}_q \frac{\big(1 + k \ln 2\big)^{\frac{q}{\gamma}}}{N_{h_k}^q}, \quad \forall k \ge \kappa, \tag{3.5.4}$$

where $\bar{C}_q = 2\left(1 \vee c^{1/\gamma}\right)^q \left(\frac{\omega_2}{\omega_3}\varepsilon_0 \vee 1\right)^q \left(1 + q\,\omega_1 \int_0^{+\infty} (\eta+1)^{q-1} e^{-\omega_3 \eta^\gamma} d\eta\right)$. We also have

$$\begin{aligned} &\mathbb{E}_f\big|\hat{f}^{(\hat{k})}(y) - f(y)\big|^q \\ &\quad \le \mathbb{E}_f\big|\hat{f}^{(\hat{k})}(y) - f(y)\big|^q \mathbb{I}_{\hat{k}\le\kappa} + \mathbb{E}_f\big|\hat{f}^{(\hat{k})}(y) - f(y)\big|^q \mathbb{I}_{\hat{k}>\kappa} \\ &\quad := R_1(f) + R_2(f). \end{aligned} \tag{3.5.5}$$

First, we control R_1. Obviously

$$\big|\hat{f}^{(\hat{k})}(y) - f(y)\big| \le \big|\hat{f}^{(\hat{k})}(y) - \hat{f}^{(\kappa)}(y)\big| + \big|\hat{f}^{(\kappa)}(y) - f(y)\big|.$$

The definition of $\hat{k}$ yields

$$\big|\hat{f}^{(\hat{k})}(y) - \hat{f}^{(\kappa)}(y)\big|\mathbb{I}_{\hat{k}\le\kappa} \le C s_n(\kappa), \quad s_n(k) = (1 + k\ln 2)^{\frac{q}{\gamma}} N_{h_k}^{-q},$$

where $C = \left(\omega_3^{-1}\frac{4\gamma + 4dq}{\gamma}\right)^{1/\gamma}$. In view of (3.5.4) we also get

$$\mathbb{E}_f\big|\hat{f}^{(\kappa)}(y) - f(y)\big|^q \le \bar{C}_q s_n(\kappa).$$

Noting that the right hand side of the obtain inequality is independent of f and taking into account the definition of κ and h^* we obtain

$$\limsup_{n\to\infty} \sup_{f\in\mathbb{H}_d(\beta,L,M)} \phi_n^{-q}(\beta) R_1(f) < \infty. \tag{3.5.6}$$

Now let us bounded from above R_2. Applying Cauchy-Schwarz inequality we have in view of Lemma 5

$$\begin{aligned} R_2(f) &= \sum_{k>\kappa}^{\mathbf{k}_n} \mathbb{E}_f\big|\hat{f}^{(k)}(y) - f(y)\big|^q I_{\left[\hat{k}=k,G\right]} \\ &\le \sum_{k>\kappa} \left(\mathbb{E}_f\big|\hat{f}^{(k)}(y) - f(y)\big|^{2q}\right)^{1/2} \sqrt{\mathbb{P}_f\big\{\hat{k}=k\big\}} \\ &= \Delta(h^*) \sum_{k>\kappa} \left(\mathbb{E}_f\big|\hat{f}^{(k)}(y) - f(y)\big|^{2q}\right)^{1/2} 2^{-(k-1)\frac{\gamma+dq}{\gamma}}, \end{aligned} \tag{3.5.7}$$

where we have put $\Delta(h^*) = J_2 \exp\big\{J_1 n (h^*)^{\gamma\beta+d}\big\}$. We obtain from (3.5.4) and (3.5.7)

$$R_2(f) \le\ J_3\ N_{h_{\max}}^{-q} \exp\big\{J_1 n (h^*)^{\gamma\beta+d}\big\}, \tag{3.5.8}$$

where

$$J_3 = J_2\ 2^{\frac{\gamma+dq}{\gamma}}\ \bar{C}_{2q}^{1/2} \sum_{s\ge 0} (1 + s\ln 2)^{\frac{q}{\gamma}} 2^{-s}.$$

It remains to note that the definition of h^* implies that

$$\limsup_{n\to\infty} \phi_n^{-q}(\beta) N_{h_{\max}}^{-q} \exp\left\{J_1 n (h^*)^{\gamma\beta+d}\right\} < \infty$$

and that the right hand side of (3.5.8) is independent of f. Thus, we have

$$\limsup_{n\to\infty} \sup_{f\in\mathbb{H}_d(\beta,L,M)} \phi_n^{-q}(\beta) R_2(f) < \infty.$$

that yields together with (3.5.5) and (3.5.6)

$$\limsup_{n\to\infty} \sup_{f\in\mathbb{H}_d(\beta,L,M)} \phi_n^{-q}(\beta) \mathbb{E}_f \left|\hat{f}^{(\hat{k})}(y) - f(y)\right|^q < \infty.$$

∎

3.5.4 Proof of Proposition 3

Auxiliary Lemmas. Set for any $a > 0$

$$Q_a = \int_{U_n \cap \left\{\|u\|_1 > a\right\}} \|u\|_1 Z_{n,\theta}^{1/m}(u) du.$$

where $U_n := N_h\big(\Theta - \theta\big)$, Θ defined in (3.1.4) and $Z_{n,\theta}$ defined in (3.2.1). Remember that $\gamma > 0$ is the constant which defined the *minimax rate of convergence* $\varphi_{n,\gamma}$ and depends on the density $g(.)$.

Lemma 6. *For any f such that $M(f) < \infty$ and $h \in \mathcal{H}_n$, if the assumption **(1)** is verified, then*

$$\mathbb{P}_f\left\{\int_{U_n} Z_{n,\theta}^{1/m}(u) du < \frac{1}{2}\delta^{D_b}\right\} < 2C_1\, \mathcal{E}_h^{c_1}\, \delta^{\tau}, \quad \forall \delta \in \left]0, 1 \wedge M\right].$$

Lemma 7. *For all $h \in \mathcal{H}_n$ and any f such that $M(f) < \infty$, if the assumption 1.2 is verified, then for any $a > s_2\, \mathcal{N}(h)$*

$$\mathbb{E}_f\big\{Q_a\big\} \;\le\; 2C_2\, a\, \mathcal{E}_h^{c_2}\, I_1\, e^{-c_3 a^{\gamma}},$$

where s_2 is defined in Section 3.2 and

$$I_1 = \int_0^{+\infty} (z+1) e^{-c_3 z^{\gamma}} dz$$

This lemmas are proved in section 3.6.

Proof of Proposition 3. The definition of $\hat{\theta}(h)$ and $\theta = \theta(f, y, h)$ implies $\forall \varepsilon > 0$

$$\begin{aligned} \mathbb{P}_f\left(N_h\big|\hat{f}^h(y) - f(y)\big| \geq \varepsilon\right) &\leq \mathbb{P}_f\left(N_h\big|\hat{\theta}_0(h) - \theta_0\big| \geq \varepsilon\right) \\ &\leq \mathbb{P}_f\left(N_h\big\|\hat{\theta}(h) - \theta\big\|_1 \geq \varepsilon\right). \end{aligned} \tag{3.5.9}$$

Let us do some remarks. First, by definition we have $\theta \in \Theta\big(M\big)$. Remind also that $\hat{\theta}(h)$ minimizes π_h defined in (3.1.6) and, therefore, the following inclusion holds since $\hat{\theta}(h) \in \Theta$.

$$\left\{N_h\big\|\hat{\theta}(h) - \theta\big\|_1 \geq \varepsilon\right\} \subseteq \left\{\inf_{N_h\|t-\theta\|_1 \geq \varepsilon} \pi_h(t) \leq \pi_h(\theta)\right\}. \tag{3.5.10}$$

Moreover,

$$\begin{aligned} \pi_h(t) &= N_h^{-1}\int_\Theta \|N_h(t-u)\|_1 \left[L_h\big(u, \mathcal{Z}_n\big)\right]^{1/m} du \\ &= N_h^{-D_b-1}\int_{U_n} \|N_h(t-\theta)-u\|_1 \left[L_h\big(\theta + uN_h^{-1}, \mathcal{Z}_n\big)\right]^{1/m} du \\ &= N_h^{-D_b-1}\left[L_h\big(\theta, \mathcal{Z}_n\big)\right]^{1/m}\int_{U_n} \|N_h(t-\theta)-u\|_1 Z_{n,\theta}^{1/m}(u)du. \end{aligned}$$

Hence, $\tau_n = N_h\big(\hat{\theta}(h) - \theta\big)$ is the minimizer of

$$\chi_n(s) = \int_{U_n} \big\|s-u\big\|_1 z_n(u)du, \quad z_n(u) = \frac{Z_{n,\theta}^{1/m}(u)}{\int_{U_n} Z_{n,\theta}^{1/m}(v)dv},$$

and we obtain from (3.5.9) and (3.5.10) for any $\varepsilon > 0$

$$\mathbb{P}_f\left\{\left\|N_h(\hat{\theta}(h) - \theta)\right\|_1 > \varepsilon\right\} \leq \mathbb{P}_f\left\{\inf_{\|s\|_1 > \varepsilon} \chi_n(s) \leq \chi_n(0)\right\}. \tag{3.5.11}$$

Let $0 < r < \varepsilon/3$, a number whose will be done later. We have

$$\chi_n(0) \leq r\int_{U_n \cap (\|u\|_1 \leq r)} z_n(u)du + \int_{U_n \cap (\|u\|_1 > r)} \|u\|_1 z_n(u)du.$$

Note also that

$$\begin{aligned} \inf_{\|s\|_1>\varepsilon} \chi_n(s) &\geq \inf_{\|s\|_1>\varepsilon}\left[\int_{U_n\cap(\|u\|_1\leq r)} \big(\|s\|_1 - \|u\|_1\big) z_n(u)du\right] \\ &\geq (\varepsilon - r)\int_{U_n\cap(\|u\|_1\leq r)} z_n(u)du. \end{aligned}$$

It yields in particular

$$\chi_n(0) - \inf_{\|s\|_1>\varepsilon} \chi_n(s) \leq -(\varepsilon - 2r) \int_{U_n \cap (\|u\|_1 \leq r)} z_n(u)du + \int_{U_n \cap (\|u\|_1 > r)} \|u\|_1 z_n(u)du.$$

Thus, $\forall r \in (0, \varepsilon/3)$

$$\begin{aligned}
&\mathbb{P}_f \left\{ \chi_n(0) - \inf_{\|s\|_1>\varepsilon} \chi_n(s) > 0 \right\} \\
&\leq \mathbb{P}_f \left\{ \int_{U_n \cap (\|u\|_1 > r)} \|u\|_1 z_n(u)du > (\varepsilon - 2r) \int_{U_n \cap (\|u\|_1 \leq r)} z_n(u)du \right\} \\
&\leq \mathbb{P}_f \left\{ \int_{U_n \cap (\|u\|_1 > r)} \|u\|_1 z_n(u)du > r/2 \right\} \\
&\quad + \mathbb{P}_f \left\{ (\varepsilon - 2r) \int_{U_n \cap (\|u\|_1 \leq r)} z_n(u)du < r/2 \right\}. \qquad (3.5.12)
\end{aligned}$$

We note that the second term in (3.5.12) can be to control by the first one whenever $0 < r < \varepsilon/3$. Indeed, putting $U_n(r) = U_n \cap (\|u\|_1 > r)$ we get

$$\begin{aligned}
&\mathbb{P}_f \left\{ (\varepsilon - 2r) \int_{U_n \cap (\|u\|_1 \leq r)} z_n(u)du < r/2 \right\} \\
&\leq \mathbb{P}_f \left\{ r \int_{U_n} Z_{n,\theta}^{1/m}(v)dv - r \int_{U_n(r)} Z_{n,\theta}^{1/m}(u)du < \frac{r}{2} \int_{U_n} Z_{n,\theta}^{1/m}(v)dv \right\} \\
&\leq \mathbb{P}_f \left\{ r \int_{U_n(r)} Z_{n,\theta}^{1/m}(u)du > \frac{r}{2} \int_{U_n} Z_{n,\theta}^{1/m}(v)dv \right\} \\
&\leq \mathbb{P}_f \left\{ \int_{U_n(r)} \|u\|_1 z_n(u)du > r/2 \right\}.
\end{aligned}$$

The last inequality and (3.5.12) yield

$$\mathbb{P}_f \left\{ \chi_n(0) - \inf_{\|s\|_1>\varepsilon} \chi_n(s) > 0 \right\} \leq 2\mathbb{P}_f \left\{ \int_{U_n(r)} \|u\|_1 z_n(u)du > r/2 \right\}.$$

Let $\nu > 0$ the parameter whose choice will be done later, then by definition of z_n and the last inequality, we obtain

$$\begin{aligned}
\mathbb{P}_f \left\{ \chi_n(0) - \inf_{\|s\|_1>\varepsilon} \chi_n(s) > 0 \right\} \;&\leq\; 2\mathbb{P}_f \left\{ \int_{U_n(r)} \|u\|_1 Z_{n,\theta}^{1/m}(u)du > r\,\nu/4 \right\} \\
&\quad + 2\mathbb{P}_f \left\{ \int_{U_n} Z_{n,\theta}^{1/m}(v)dv < \nu/2 \right\}. \qquad (3.5.13)
\end{aligned}$$

In view of Lemma 7 and the Markov inequality, choosing $r = \varepsilon/4 \geq s_2$ we get

$$\mathbb{P}_f \left\{ \int_{U_n \cap (\|u\|_1 > \varepsilon/4)} \|u\|_1 Z_{n,\theta}^{1/m}(u) du > \frac{\varepsilon \nu}{16} \right\} \leq 8 C_2 \, I_1 \, \mathcal{E}_h^{c_2} \, \nu^{-1} \, e^{-c_3 2^{-2\gamma} \varepsilon^\gamma}. \tag{3.5.14}$$

Applying Lemma 6, we have for $\nu \leq M$

$$\mathbb{P}_f \left\{ \int_{U_n} Z_{n,\theta}^{1/m}(v) dv < \nu/2 \right\} \leq 2 C_1 \, \mathcal{E}_h^{c_1} \nu^{\frac{\tau}{D_b}} \tag{3.5.15}$$

In view of (3.5.13), (3.5.15), we establish that

$$\mathbb{P}_f \left\{ \chi_n(0) - \inf_{\|s\|_1 > \varepsilon} \chi_n(s) > 0 \right\} \leq 16 C_2 I_1 \, \mathcal{E}_h^{c_2} \, \nu^{-1} e^{-c_3 2^{-2\gamma} \varepsilon^\gamma} + 4 C_1 \, \mathcal{E}_h^{c_1} \mathcal{E}_h^{c_1} \nu^{\frac{\tau}{D_b}}.$$

Choosing $\nu = e^{-c_3 2^{-2\gamma-1} \varepsilon^\gamma}$, we obtain

$$\mathbb{P}_f \left\{ \chi_n(0) - \inf_{\|s\|_1 > \varepsilon} \chi_n(s) > 0 \right\} \leq 16 C_2 I_1 \, \mathcal{E}_h^{c_2} \, e^{-c_3 2^{-2\gamma-1} \varepsilon^\gamma} + 4 C_1 \, \mathcal{E}_h^{c_1} e^{-c_3 \frac{2^{-2\gamma-1} \tau}{D_b} \varepsilon^\gamma}.$$

The assertion of the theorem follows now from (3.5.9), (3.5.13) and the definitions of ω_1, ω_2, ω_3. ∎

3.6 Appendix

In this part, we consider $nh^d \in \mathbb{N}^*$, in order to simplify the proofs.

Proof of Lemma 1. **1.** By definition of model 3.4.1, we have $\forall u \in U_n$

$$Z_{n,\theta}(u) = \prod_{X_i \in V_h(y)} \exp \left\{ -\frac{1}{2} \big(Y_i - f_{\theta + u \, N_h^{-1}}(X_i) \big)^2 + \frac{1}{2} \big(Y_i - f_\theta(X_i) \big)^2 \right\}.$$

Let us give the notation

$$\Sigma = \sum_{X_i \in V_h(y)} \frac{1}{4} \big(Y_i - f_{\theta + u \, N_h^{-1}}(X_i) \big)^2 - \frac{1}{4} \big(Y_i - f_\theta(X_i) \big)^2.$$

Applying $1 - e^{-\rho} \leq \rho$, $\rho \in \mathbb{R}$ and Cauchy-Schwarz inequality, we obtain for any $u : \|u\|_1 < 1$

$$\begin{aligned} \mathbb{E}_f \big(1 - Z_{n,\theta}^{1/2}(u) \big)_+ &= \mathbb{E}_f \big(1 - e^{-\Sigma} \big) \, \mathbb{I}_{[0,+\infty[}(\Sigma) \\ &\leq \mathbb{E}_f (\Sigma) \, \mathbb{I}_{[0,+\infty[}(\Sigma) \\ &\leq 2 \mathcal{E}_h \|u\|_1^2. \end{aligned}$$

Using the last inequality and define $\Gamma_\delta = [0,\delta]^{D_b}$, we establish

$$\int_{[0,\delta]^{D_b}} \mathbb{E}_f \left[1 - Z_{n,\theta}^{1/2}(u)\right]_+ du \leq 2\mathcal{E}_h \delta^{D_b+2}, \quad \forall \delta < 1.$$

2. Using $\mathbb{E}e^{\lambda\varepsilon} = e^{\lambda^2/2}$, with $\varepsilon \sim \mathcal{N}(0,1)$ and $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}_f Z_{n,\theta}^{1/2}(u) = \exp\left\{-\frac{1}{8}\sum_{X_i \in V_h(y)} f^2_{u\, N_h^{-1}}(X_i) + \frac{1}{2}\sum_{X_i \in V_h(y)} f_{u\, N_h^{-1}}(X_i)\big(f(X_i) - f_\theta(X_i)\big)\right\}. \tag{3.6.1}$$

Remark that $f_{u\, N_h^{-1}}(X_i) \leq \|u\|_1 N_h^{-1}$, $f(X_i) - f_\theta(X_i) \leq b_h$ and

$$\sum_{X_i \in V_h(y)} f^2_{u\, N_h^{-1}}(X_i) = u^\top \mathcal{M}_{nh}(y)\, u.$$

In view of Lemma 9 in Chapter 4, $\exists \lambda > 0$ such that $\lambda_n(h) \geq \lambda$ and $u^\top \mathcal{M}_{nh}(y) u \geq \lambda_n(h) u^\top u$. Applying (3.6.1), we have

$$\mathbb{E}_f Z_{n,\theta}^{1/2}(u) \leq \exp\left\{-\frac{\lambda_n(h)}{8}\|u\|_2^2 + \frac{\|u\|_1}{2}\mathcal{N}(h)\right\},$$

By equivalence between the ℓ_1-norm and ℓ_2-norm, we obtain

$$\begin{aligned}
\mathbb{E}_f Z_{n,\theta}^{1/2}(u) &\leq \exp\left\{-\frac{D_b\lambda_n(h)}{8}\|u\|_1^2 + \frac{\|u\|_1}{2}\mathcal{N}(h)\right\} \\
&\leq \exp\left\{-\frac{D_b\lambda_n(h)}{16}\,\|u\|_1^2\right\}, \quad \forall u \in U_n : \ \|u\|_1 \geq \frac{8}{D_b\,\lambda_n(h)}\mathcal{N}(h).
\end{aligned}$$

∎

Proof of Lemma 2. **1.** Remember that we have $\forall u \in U_n = N_h\big(\Theta\big(A, M\big) - \theta\big)$

$$Z_{n,\theta}(u) = \prod_{X_i \in V_y(h)} \left(1 + \frac{f_{uN_h^{-1}}(X_i)}{f_\theta(X_i)}\right)^{Y_i} \exp\left\{-f_{uN_h^{-1}}(X_i)\right\}, \quad N_h = \big(nh^d\big)^{1/2}.$$

Using the Cauchy-Schwarz inequality, we obtain for $u \in U_n$: $\|u\|_1 < 1$

$$\mathbb{E}_f\big(1 - Z_{n,\theta}^{1/2}(u)\big)_+ \leq \Big(\mathbb{E}_f\big|1 - Z_{n,\theta}^{1/2}(u)\big|^2\Big)^{1/2} = \Big(\mathbb{E}_f\big(1 + Z_{n,\theta}(u) - 2Z_{n,\theta}^{1/2}(u)\big)\Big)^{1/2}. \tag{3.6.2}$$

It is easy to show that for $n \geq n_0$

$$\mathbb{E}_f Z_{n,\theta}(u) \leq 1 + \frac{A + \mathcal{N}(h)}{A}\|u\|_1.$$

So, using (3.6.2) we have

$$\mathbb{E}_f\big(1-Z_{n,\theta}^{1/2}(u)\big)_+ \le \left(\frac{A+\mathcal{N}(h)}{A}\|u\|_1 + 2\big(1-\mathbb{E}_f Z_{n,\theta}^{1/2}(u)\big)\right)^{1/2}. \tag{3.6.3}$$

It remains to calculate the term

$$\begin{aligned}
\mathbb{E}_f Z_{n,\theta}^{1/2}(u) &= \prod_{X_i\in V_h(y)} \mathbb{E}_f\left(1+\frac{f_{uN_h^{-1}}(X_i)}{f_\theta(X_i)}\right)^{Y_i/2} \exp\Big\{-f_{uN_h^{-1}}(X_i)/2\Big\} \\
&= \prod_{X_i\in V_h(y)} \sum_{k=0}^{\infty} \frac{\big(f(X_i)\big)^k}{k!}\left(1+\frac{f_{uN_h^{-1}}(X_i)}{f_\theta(X_i)}\right)^{k/2} \exp\left\{-f(X_i)-\frac{1}{2}f_{uN_h^{-1}}(X_i)\right\} \\
&= \prod_{X_i\in V_h(y)} \exp\left\{-f(X_i)-\frac{1}{2}f_{uN_h^{-1}}(X_i)+f(X_i)\left(1+\frac{f_{uN_h^{-1}}(X_i)}{f_\theta(X_i)}\right)^{1/2}\right\} \qquad (3.6.4)
\end{aligned}$$

Using the Taylor expansion series, for $n \ge n_0$ we get

$$\left(1+\frac{f_{uN_h^{-1}}(X_i)}{f_\theta(X_i)}\right)^{1/2} = 1+\frac{1}{2}\frac{f_{uN_h^{-1}}(X_i)}{f_\theta(X_i)} - \frac{1}{4}\left(\frac{f_{uN_h^{-1}}(X_i)}{f_\theta(X_i)}\right)^2 + O\left[\left(\frac{f_{uN_h^{-1}}(X_i)}{f_\theta(X_i)}\right)^3\right],$$

where $O(\rho^3) \xrightarrow{\rho\to 0} 0$. Applying last equality and (3.6.4), we have

$$\begin{aligned}
\mathbb{E}_f Z_{n,\theta}^{1/2}(u) &= \prod_{X_i\in V_h(y)} \exp\left\{-\frac{1}{2}f_{uN_h^{-1}}(X_i)\frac{f_\theta(X_i)-f(X_i)}{f_\theta(X_i)} - \frac{1}{4}\left(f_{uN_h^{-1}}(X_i)\frac{f(X_i)}{f_\theta(X_i)}\right)^2\right. \\
&\qquad \left. + f(X_i)\, O\left[\left(\frac{f_{uN_h^{-1}}(X_i)}{f_\theta(X_i)}\right)^3\right]\right\}. \qquad (3.6.5)
\end{aligned}$$

Using the inequality $1-e^{-\rho} \le \rho$, $\rho \ge 0$, (3.6.3), (3.6.4) and (3.6.5), we have for $n \ge n_0$,

$$\mathbb{E}_f\big(1-Z_{n,\theta}^{1/2}(u)\big)_+ \le \sqrt{\frac{2+M+2A}{2A}}\mathcal{E}_h^{1/2}\|u\|_1^{1/2}, \quad u\in[0,\delta]^{D_b}, \quad \delta<1.$$

By integration,the first assertion of the lemma is proved.

2. Using (3.6.5) we have for $n \ge n_0$ and $\forall u \in U_n$: $\|u\|_1 \ge \frac{4M^2}{\lambda_n(h)}\mathcal{N}(h)$, in view of Lemma 9 in Chapter 4,

$$\mathbb{E}_f Z_{n,\theta}^{1/2}(u) \le e^1 \exp\left\{-\frac{\lambda_n(h)A}{8M^2}\|u\|_1^2\right\},$$

where $\lambda_n(h)$ is defined in Section 3.4. ∎

Proof of Lemma 3. Remember that we have $\forall u \in U_n = N_h\big(\Theta\big(M\big) - \theta\big)$

$$\begin{aligned} Z_{n,\theta}(u) &= \frac{L_h\big(\theta + u\, N_h^{-1}, \mathcal{Z}_n\big)}{L_h\big(\theta, \mathcal{Z}_n\big)} \\ &= \prod_{X_i \in V_y(h)} \exp\Big\{-\big|Y_i - f_{\theta+uN_h^{-1}}(X_i)\big|^\alpha + \big|Y_i - f_\theta(X_i)\big|^\alpha\Big\}, \quad N_h = \big(nh^d\big)^{1/(1+2\alpha)}. \end{aligned}$$

It is easy to see that for $m \geq 1$

$$\mathbb{E}_f Z_{n,\theta}^{1/m}(u) = \prod_{X_i \in V_h(y)} \int_{\mathbb{R}} C(\alpha) \exp\left\{-\frac{1}{m}|y - b_i - v_i|^\alpha + \frac{1}{m}|y - b_i|^\alpha - |y|^\alpha\right\},$$

where $b_i = f_\theta(X_i) - f(X_i)$ and $v_i = f_{uN_h^{-1}}(X_i)$. Remark that if we choose the approximation polynomial f_θ such that $f_\theta \geqq f$ (for construction, see Chapter 4), then $b_i \geq 0, \forall i$. Without loss of generality we choose $v_i \geq 0$. Introduce

$$\mathcal{G}_m(v) = \int_{\mathbb{R}} C(\alpha) \exp\left\{-\frac{1}{m}|y - b - v|^\alpha + \frac{1}{m}|y - b|^\alpha - |y|^\alpha\right\}, \quad v \geq 0, \quad b \geq 0.$$

We will prove that $\mathcal{G}_m(v) \leq 1 - T_1 v^{1+2\alpha}$ for $m > 1$. Indeed, we use standard schema such that

$$\mathcal{G}_m(v) = \mathcal{G}_m(0) + \int_0^v \mathcal{G}_m'(\tilde{v}) d\tilde{v}. \tag{3.6.6}$$

It remains to prove that $\mathcal{G}_m'(\tilde{v}) \leq -T_2 \tilde{v}^{2\alpha}$. For it, we decompose the function $\mathcal{G}_m$ of the following way

$$\begin{aligned} \mathcal{G}_m(v) &= \int_v^\infty C(\alpha) \exp\left\{-\frac{1}{m}(y - v)^\alpha + \frac{1}{m}y^\alpha - (y + b)^\alpha\right\} dy \\ &+ \int_0^v C(\alpha) \exp\left\{-\frac{1}{m}(v - y)^\alpha + \frac{1}{m}y^\alpha - (y + b)^\alpha\right\} dy \\ &+ \int_{-b}^0 C(\alpha) \exp\left\{-\frac{1}{m}(v - y)^\alpha + \frac{1}{m}(-y)^\alpha - (y + b)^\alpha\right\} dy \\ &+ \int_{-\infty}^{-b} C(\alpha) \exp\left\{-\frac{1}{m}(v - y)^\alpha + \frac{1}{m}(-y)^\alpha - (y + b)^\alpha\right\} dy \end{aligned}$$

The derivative of $\mathcal{G}_m$ can be formulate like

$$\begin{aligned}
\mathcal{G}'_m(v) &= -C(\alpha)\exp\left\{\frac{v^\alpha}{m}-(v+b)^\alpha\right\}+C(\alpha)\exp\left\{\frac{v^\alpha}{m}-(v+b)^\alpha\right\}\\
&+\int_v^\infty C(\alpha)\frac{\alpha}{m}(y-v)^{\alpha-1}\exp\left\{-\frac{1}{m}(y-v)^\alpha+\frac{1}{m}y^\alpha-(y+b)^\alpha\right\}dy\\
&-\int_0^v C(\alpha)\frac{\alpha}{m}(v-y)^{\alpha-1}\exp\left\{-\frac{1}{m}(v-y)^\alpha+\frac{1}{m}y^\alpha-(y+b)^\alpha\right\}dy\\
&-\int_{-b}^0 C(\alpha)\frac{\alpha}{m}(v-y)^{\alpha-1}\exp\left\{-\frac{1}{m}(v-y)^\alpha+\frac{1}{m}(-y)^\alpha-(y+b)^\alpha\right\}dy\\
&-\int_{-\infty}^{-b} C(\alpha)\frac{\alpha}{m}(v-y)^{\alpha-1}\exp\left\{-\frac{1}{m}(v-y)^\alpha+\frac{1}{m}(-y)^\alpha-(y+b)^\alpha\right\}dy.
\end{aligned}$$

By changing variables and summation of integrals, we obtain

$$\begin{aligned}
\mathcal{G}'_m(v) &= \int_{v+b}^\infty C(\alpha)\frac{\alpha}{m}z^{\alpha-1}\exp\left\{-\frac{1}{m}z^\alpha\right\}\left[\exp\left\{\frac{1}{m}(z+v)^\alpha-(z+v+b)^\alpha\right\}\right.\\
&\qquad\left.-\exp\left\{\frac{1}{m}(z-v)^\alpha-(z-v-b)^\alpha\right\}\right]dz \qquad (3.6.7)\\
&+\int_v^{v+b} C(\alpha)\frac{\alpha}{m}z^{\alpha-1}\exp\left\{-\frac{1}{m}z^\alpha\right\}\left[\exp\left\{\frac{1}{m}(z+v)^\alpha-(z+v+b)^\alpha\right\}\right.\\
&\qquad\left.-\exp\left\{\frac{1}{m}(z-v)^\alpha-(v-z+b)^\alpha\right\}\right]dz \qquad (3.6.8)\\
&+\int_0^v C(\alpha)\frac{\alpha}{m}z^{\alpha-1}\exp\left\{-\frac{1}{m}z^\alpha\right\}\left[\exp\left\{\frac{1}{m}(z+v)^\alpha-(z+v+b)^\alpha\right\}\right.\\
&\qquad\left.-\exp\left\{\frac{1}{m}(v-z)^\alpha-(v-z+b)^\alpha\right\}\right]dz. \qquad (3.6.9)
\end{aligned}$$

By expansion Taylor series of order 1 and for any $n \geq n_0$, the term (3.6.9) can be controlled by

$$\begin{aligned}
&\int_0^v C(\alpha)\frac{\alpha}{m}z^{\alpha-1}\exp\left\{-\frac{1}{m}z^\alpha\right\}\left(\frac{b^\alpha}{m}-\frac{m-1}{m}(v-z+b)^\alpha-\frac{m-1}{m}(z+v+b)^\alpha+O\big(b^{2\alpha}\big)\right)dz\\
&\quad\leq\big(1+C(\alpha)\big)\frac{v^\alpha b^\alpha}{m^2}.
\end{aligned}$$

The second term (3.6.8) can be bounded by $\big(1+C(\alpha)\big)\frac{v^\alpha b^\alpha}{m^2}$. We note that the first term is the integral of a negative function, so we can increase it by

$$\begin{aligned}
&\int_{v+b}^{2v+b} C(\alpha)\frac{\alpha}{m}z^{\alpha-1}\exp\left\{-\frac{1}{m}z^\alpha\right\}\left(-\frac{m-1}{m}(z-v-b)^\alpha-\frac{m-1}{m}(z+v+b)^\alpha+O\big(v^{2\alpha}\big)\right)dz\\
&\quad\le -\int_{v+b}^{2v+b} C(\alpha)\frac{\alpha}{m}z^{\alpha-1}\exp\left\{-\frac{1}{m}z^\alpha\right\}\frac{m-1}{m}(2z)^\alpha dz+O\big(v^{3\alpha}\big)\\
&\quad\le -v^{2\alpha}\,2^{\alpha-2}\,C(\alpha)\frac{m-1}{m^2}\exp\left\{-\frac{1}{m}v^\alpha\right\}.
\end{aligned}$$

So we obtain

$$\mathcal{G}_m'(v)\le 2\big(1+C(\alpha)\big)\frac{v^\alpha b^\alpha}{m^2}-2^{\alpha-2}\,C(\alpha)\frac{m-1}{m^2}\exp\left\{-\frac{1}{m}v^\alpha\right\}v^{2\alpha},$$

Using (3.6.6), the last inequality implies that $\forall m>1$

$$\mathcal{G}_m(v)\le 1+2\big(1+C(\alpha)\big)\frac{b^\alpha|v|^{1+\alpha}}{(1+\alpha)m^2}-2^{\alpha-2}\,C(\alpha)\frac{m-1}{(1+2\alpha)m^2}\exp\left\{-\frac{1}{m}v^\alpha\right\}|v|^{1+2\alpha},\tag{3.6.10}$$

with $\mathcal{G}_m(0)=1$. We can also show that it exists positive constants T_3, T_4 such that

$$\mathcal{G}_1(v)\le 1+T_3|v|^{1+2\alpha}+T_4|v|^{1+\alpha}b^\alpha,\quad v\in\mathbb{R}.\tag{3.6.11}$$

1. Let us prove the first assumption,

$$\mathbb{E}_f\big(1-Z_{n,\theta}^{1/2}(u)\big)_+\le\mathbb{E}_f\big|1-Z_{n,\theta}^{1/2}(u)\big|=\mathbb{E}_f\frac{1}{L_h^{1/2}\big(\theta,\mathcal{Z}_n\big)}\Big|L_h^{1/2}\big(\theta,\mathcal{Z}_n\big)-L_h^{1/2}\big(\theta+u,\mathcal{Z}_n\big)\Big|.$$

Let us note $p_h(x,f)=\prod_{X_i\in V_h(y)}g_\alpha\big(x_i,f(X_i)\big)$, where $x\in\mathbb{R}^{nh^d}$. By integration and Cauchy-Schwarz inequality, we obtain that

$$\begin{aligned}
\mathbb{E}_f\big(1-Z_{n,\theta}^{1/2}(u)\big)_+ &\le \int_{\mathbb{R}^{nh^d}}\Big|p_h^{1/2}\big(x,f_{\theta+u\,N_h^{-1}}\big)-p_h^{1/2}\big(x,f_\theta\big)\Big|\frac{p_h\big(x,f\big)}{p_h^{1/2}\big(x,f_\theta\big)}\,dx\\
&\le \left(\int_{\mathbb{R}^{nh^d}}\Big|p_h^{1/2}\big(x,f_{\theta+u\,N_h^{-1}}\big)-p_h^{1/2}\big(x,f_\theta\big)\Big|^2dx\int_{\mathbb{R}^{nh^d}}\frac{p_h^2\big(x,f\big)}{p_h\big(x,f_\theta\big)}\,dx\right)^{1/2}.
\end{aligned}\tag{3.6.12}$$

Introduce

$$H^2\left(\mathbb{P}_{f_{\theta+u\,N_h^{-1}}},\mathbb{P}_{f_\theta}\right)=\int_{\mathbb{R}^{nh^d}}\Big|p_h^{1/2}\big(x,f_{\theta+u\,N_h^{-1}}\big)-p_h^{1/2}\big(x,f_\theta\big)\Big|^2dx,$$

where $H(.,.)$ is the Hellinger distance between probability measures. More details are given in Tsybakov [2008] (Chapter 2). Remark that $\mathbb{P}_{f_\theta}$ is a product measure $\mathbb{P}_{f_\theta} = \otimes_{X_i \in V_h(y)} \mathbb{P}^i_{f_\theta}$ where $\mathbb{P}^i_{f_\theta} = \mathbb{P}_{f_\theta(X_i)}$ the probability law of Y_i. By property of Hellinger distance, we obtain

$$H^2\left(\mathbb{P}_{f_{\theta+u\,N_h^{-1}}}, \mathbb{P}_{f_\theta}\right) = 2\left[1 - \prod_{X_i \in V_h(y)} \left(1 - \frac{1}{2} H^2\left(\mathbb{P}^i_{f_{\theta+u\,N_h^{-1}}}, \mathbb{P}^i_{f_\theta}\right)\right)\right], \tag{3.6.13}$$

where

$$\begin{aligned} H^2\left(\mathbb{P}^i_{f_{\theta+u\,N_h^{-1}}}, \mathbb{P}^i_{f_\theta}\right) &= \int_{\mathbb{R}} \left| g_\alpha^{1/2}\big(x - f_{\theta+u\,N_h^{-1}}(X_i)\big) - g_\alpha^{1/2}\big(x - f_\theta(X_i)\big)\right|^2 dx, \\ &= \int_{\mathbb{R}} \left| g_\alpha^{1/2}\big(x - f_{u\,N_h^{-1}}(X_i)\big) - g_\alpha^{1/2}(x)\right|^2 dx, \quad i \,:\, X_i \in V_h(y). \end{aligned}$$

Has'minskii and Ibragimov [1981] showed that g_α possess singularities of order 2α for $\alpha < 1/2$. In particular they gave the following result on g_α:

$$\int_{\mathbb{R}} \left| g_\alpha^{1/2}(x-\eta) - g_\alpha^{1/2}(x)\right|^2 dx = O\big(\eta^{1+2\alpha}\big), \quad \eta \to 0. \tag{3.6.14}$$

For more details, see Has'minskii and Ibragimov [1981] (Chapter 6, section 1, Theorem 1.1). Using (3.6.12), (3.6.13) and 3.6.14, then it exists a constant $\mathcal{C}$ such that for $n \geq n_0$

$$\mathbb{E}_f\big(1 - Z^{1/2}_{n,\theta}(u)\big)_+ \leq \mathcal{C}\|u\|_1^{\alpha+1/2} \left(\int_{\mathbb{R}^{nhd}} \frac{p_h^2(x,f)}{p_h(x,f_\theta)}\, dx\right)^{1/2}. \tag{3.6.15}$$

In order to show the assumption 1.1, it remains to prove the following assertion

$$\left(\int_{\mathbb{R}^{nhd}} \frac{p_h^2(x,f)}{p_h(x,f_\theta)}\, dx\right)^{1/2} \leq \exp\left\{\frac{T_3+T_4}{2}\mathcal{N}(h)^{1+2\alpha}\right\}. \tag{3.6.16}$$

For it, we note that

$$\int_{\mathbb{R}^{nhd}} \frac{p_h^2(x,f)}{p_h(x,f_\theta)}\, dx = \prod_{X_i \in V_h(y)} \mathcal{G}_1(-b_i).$$

So, applying (3.6.11) with $v = -b_i$, we obtain (3.6.16). Using (3.6.15), the first assertion of the lemma is proved.

2. Using (3.6.10) with $m = 2$ and $v = u_i$, we obtain

$$\mathbb{E}_f Z^{1/2}_{n,\theta}(u) \leq \exp\left\{\frac{1+C(\alpha)}{2(1+\alpha)}\mathcal{N}(h) - \frac{2^{\alpha-4}\,C(\alpha)}{(1+2\alpha)\exp\left\{\frac{1}{2}(2M)^\alpha\right\}} \sum_{X_i \in V_h(y)} \frac{|f_u(X_i)|^{1+2\alpha}}{nh^d}\right\}.$$

Remark that $\sum_{X_i\in V_h(y)} \frac{|f_u(X_i)|^{1+2\alpha}}{nh^d} \geq \sum_{X_i\in V_h(y)} \frac{|f_u(X_i)|^2}{nh^d\|u\|_1^{1-2\alpha}}$ and

$$\sum_{X_i\in V_h(y)} f^2_{u\, N_h^{-1}}(X_i) = u^\top \mathcal{M}_{nh}(y)\, u.$$

In view of Lemma 9 (in Chapter 4) and last inequality, we have

$$\mathbb{E}_f Z^{1/2}_{n,\theta}(u) \leq \exp\left\{\frac{1+C(\alpha)}{2(1+\alpha)}\mathcal{N}(h)\right\} \times \exp\left\{-\frac{\lambda_n(h)2^{\alpha-4}\,C(\alpha)}{(1+2\alpha)\exp\left\{\frac{1}{2}(2M)^\alpha\right\}}\|u\|_1^{1+2\alpha}\right\},$$

where $\lambda_n(h)$ is defined in Section 3.4. ∎

Proof of Lemma 5. Note that for any $k \geq \kappa+1$ and by definition of $\hat{k}$ (3.3.1)

$$\left\{\hat{k}=k\right\} = \cup_{l\geq k}\left\{\left|\hat{f}^{(k-1)}(y) - \hat{f}^{(l)}(y)\right| > S_n(l)\right\}.$$

Note that $S_n(l)$ is monotonically increasing in l and, therefore,

$$\begin{aligned}\left\{\hat{k}=k\right\} \;\subseteq\; & \left\{\left|\hat{f}^{(k-1)}(y) - f(y)\right| > 2^{-1}S_n(k-1)\right\} \\ & \cup\left[\cup_{l\geq k}\left\{\left|\hat{f}^{(l)}(y) - f(y)\right| > 2^{-1}S_n(l)\right\}\right].\end{aligned}$$

We come to the following inequality: for any $k \geq \kappa+1$

$$\begin{aligned}\mathbb{P}_f\big(\hat{k}=k\big) \;\leq\; & \mathbb{P}_f\left\{\left|\hat{f}^{(k-1)}(y) - \hat{f}(y)\right| > 2^{-1}S_n(k-1)\right\} \\ & + \sum_{l\geq k}\mathbb{P}_f\left\{\left|\hat{f}^{(l)}(y) - f(y)\right| > 2^{-1}S_n(l)\right\}. \qquad (3.6.17)\end{aligned}$$

Note that the definition of $S_n(l)$ yields

$$N_{h_l}\, S_n(l) = \left(s_2^\gamma \vee \frac{2^{2\gamma+1}(2\gamma+2dq)}{c_3\gamma\left(1\wedge\tau\, D_b^{-1}\right)}\right)^{\frac{1}{\gamma}}\left[1+\ln\left(h_{\max}/h_l\right)\right]^{1/\gamma}.$$

Thus, applying Proposition 3 with $J_1 = Ld\omega_2$ and $\varepsilon = N_{h_l}\, S_n(l)$, we obtain $\forall l \geq k-1$

$$\begin{aligned}\mathbb{P}_f\left\{\left|\hat{f}^{(l)}(y) - f(y)\right| > 2^{-1}S_n(l)\right\} \;&\leq\; \omega_1\mathcal{E}^{\omega_2}_{h_l}\left[h_{\max}/h_l\right]^{-\frac{2\gamma+2dq}{\gamma}} \\ &=\; \omega_1\mathcal{E}^{\omega_2}_{h_l}2^{-l\frac{2\gamma+2dq}{\gamma}}. \qquad (3.6.18)\end{aligned}$$

Note that $b_{h_l} \leq Ld\, h_l^\beta$ and $h_l \leq h_\kappa \leq h^*$ since $f \in \mathbb{H}_d(\beta, L, M)$ and, therefore,

$$\mathcal{N}(h_l) \leq L\, d\, n(h_l)^{\gamma\beta+d} \leq L\, d\, n(h^*)^{\gamma\beta+d}, \quad \forall l \geq k-1. \tag{3.6.19}$$

Here we have also used that $k \geq \kappa + 1$. We obtain from (3.6.17), (3.6.18) and (3.6.19) that $k \geq \kappa + 1$

$$\mathbb{P}_f\big(\hat{k} = k\big) \leq J_2 \exp\big\{J_1 n(h^*)^{\gamma\beta+d}\big\} 2^{-(k-1)\frac{2\gamma+2dq}{\gamma}},$$

where $J_2 = \omega_1 \left(1 - 2^{-\frac{2\gamma+2dq}{\gamma}}\right)^{-1}$. ∎

Proof of Lemma 6. Applying Markov inequality, Fubini theorem and assumption 1.1

$$\begin{aligned}
&\mathbb{P}_f\left\{\int_{U_n} Z_{n,\theta}^{1/m}(u)du < \frac{1}{2}\delta^{D_b}, G\right\} \\
&\quad = \mathbb{P}_f\left\{\int_{\Gamma_\delta} \big(Z_{n,\theta}^{1/m}(u) - 1\big)du < -\frac{1}{2}\delta^{D_b}, \right\} \\
&\quad \leq \mathbb{P}_f\left\{\int_{\Gamma_\delta} \big(1 - Z_{n,\theta}^{1/m}(u)\big)_+ du > \frac{1}{2}\delta^{D_b}\right\} \\
&\quad \leq 2\delta^{-D_b} \int_{\Gamma_\delta} \mathbb{E}_f\big(1 - Z_{n,\theta}^{1/m}(u)\big)_+ du \\
&\quad \leq 2C_1 \mathcal{E}_h^{c_1} \delta^\tau.
\end{aligned}$$

∎

Proof of Lemma 7. Applying the assumption 1.2 we obtain

$$\mathbb{E}_f \int_{U_n \cap (\|u\|_1 > a)} \|u\|_1 Z_{n,\theta}^{1/m}(u)\, du \leq C_2 \mathcal{E}_h^{c_2} \int_{U_n \cap (\|u\|_1 > a)} \|u\|_1 \exp\{-c_3 \|u\|_1^\gamma\}\, du.$$

Simplifying the last inequality, we have

$$\mathbb{E}_f\big\{Q_a\big\} \leq 2C_2\, a\, \mathcal{E}_h^{c_2}\, \exp\{-c_3 a^\gamma\} \int_0^{+\infty} (z+1) \exp\{-c_3 z^\gamma\}\, dz.$$

∎

Chapter 4

Locally Bayesian Approach for Multiplicative Uniform Regression

Ce chapitre correspond à l'article Chichignoud [2010a]. Dans cet article, nous présentons *l'estimateur Bayésien* sous une forme différente que celle dans le chapitre précédent. La démonstration des grandes déviations est plus complexe et longue que dans le chapitre précédent. Pour le lecteur ce chapitre contient la preuve des hypothèses 1 pour le modèle de régression avec un bruit multiplicatif uniforme (Voir lemme 8) , ainsi que la borne inférieure du risque minimax (Section 4.2) et l'optimalité de la vitesse adaptative (Section 4.3) pour ce modèle.

4.1 Introduction

Let statistical experiment be generated by the couples of observations $Y^{(n)} = (X_i, Y_i)_{i=1,\dots n}, n \in \mathbb{N}^*$ where (X_i, Y_i) satisfies the equation

$$Y_i = f(X_i) \times U_i, \quad i = 1, \dots, n. \tag{4.1.1}$$

Here $f : [0,1]^d \to \mathbb{R}$ is unknown function and we are interested in estimating f at a given point $y \in [0,1]^d$ from observation $Y^{(n)}$.

The random variables (noise) $(U_i)_{i\in 1,\dots,n}$ are supposed to be independent and uniformly distributed on $[0,1]$.

The design points $(X_i)_{i\in 1,\dots,n}$ are deterministic and without loss of generality we will assume that

$$X_i \in \left\{1/n^{1/d}, 2/n^{1/d}, \dots, 1\right\}^d, \ \ i = 1, \dots, n.$$

Along the paper the unknown function f is supposed to be smooth, in particular, it belongs to the Holder ball of functions $\mathbb{H}_d(\beta, L, M)$ (see Definition 13 below). Here $\beta > 0$ is the smoothness of f, M is the upper bound of f and its partial derivatives and $L > 0$ is Lipschitz constant.

Moreover, we will consider only the functions f separated away from zero by some positive constant. Thus, from now on we will suppose that there exists $0 < A < M$ such that $f \in \mathbb{H}_d(\beta, L, M, A)$, where

$$\mathbb{H}_d(\beta, L, M, A) = \left\{ g \in \mathbb{H}_d(\beta, L, M) : \inf_{x \in [0,1]^d} g(x) \geq A \right\}.$$

Motivation. The multiplicative regression model is quite popular in various domains of applications, in particular, in volatility estimation Härlde and Tsybakov [1997] or in noise speckle imaging Aubert and Aujol [2008]. Another source of the interest to multiplicative regression is so-called *nonparametric frontier model* (see Simar and Wilson [2000]), where the reconstruction of the regression function f can be viewed as the estimation a production set $\mathcal{P}$. Indeed, $Y_i \leq f(X_i)$, $\forall i$, and, therefore, the estimation of f is reduced to finding the upper boundary of $\mathcal{P}$. In this context one can also cite Korostelëv and Tsybakov [1993] dealing with the estimation of function's support.

The theoretical interest to the multiplicative regression model (4.1.1) with discontinuous noise is dictated by the following fact. The typical approach to the study of the models with multiplicative noise consists in their transformation into the model with an additive noise and in the application, after that, the linear smoothing technique, based on standard methods like kernel smoothing, local polynomials etc. Let us illustrate the latter approach by the consideration of one of the most popular non-parametric model namely multiplicative gaussian regression

$$Y_i = \sigma(X_i)\xi_i, \quad i = 1, \dots, n. \tag{4.1.2}$$

Here ξ_i, $i = 1, \dots, n$ are i.i.d. standard gaussian random variables and the goal is to estimate the variance $\sigma^2(\cdot)$.

Putting $Y_i' = Y_i^2$ and $\eta_i = \xi_i^2 - 1$ one can transform the model (4.1.2) into the heteroscedastic additive regression:

$$Y_i' = \sigma^2(X_i) + \sigma^2(X_i)\eta_i, \quad i = 1, \dots, n, \tag{4.1.3}$$

where, obviously, $\mathbb{E}\eta_i = 0$. Applying any of the linear methods mentioned above to the estimation of $\sigma^2(\cdot)$ one can construct an estimator whose estimation accuracy is given by $n^{-\frac{\beta}{2\beta+d}}$ and which is optimal in minimax sense (See Definition 14). The latter result is proved under assumptions on $\sigma^2(\cdot)$ which are similar to the assumption imposed on the function $f(\cdot)$. In particular, β denotes the regularity of the function $\sigma^2(\cdot)$. The same result can be obtained for any noise variables ξ_i with known, continuously differentiable density, possessing sufficiently many moments.

The situation changes dramatically when one considers the noise with discontinuous distribution density. Although, the transformation of the original multiplicative model to the additive one is still possible, in particular, the model (4.1.1) can be rewritten as

$$Y_i' = f(X_i) + f(X_i)\eta_i, \quad Y_i' = 2Y_i, \;\; \eta_i = 2u_i - 1, \quad i = 1, \dots, n,$$

the linear methods are not optimal anymore. As it is proved in Theorem 10 the optimal accuracy is given by $n^{-\frac{\beta}{\beta+d}}$. To achieve this rate the non-linear estimation procedure, based on locally bayesian approach, is proposed in Section 4.2.

The interesting feature of the model is that the uniform distribution leads to an improved estimation rate, it is always the case with a discontinuous density of the noise (see Has'minskii and Ibragimov [1981], Chapter 6). The parameter of a uniform $[0,\theta]$ distribution is easier to estimate. Indeed, it is well known that we can use the maximum likelihood estimator based on the maximum of observations (nonlinear). For instance, if f is known to be Lipshitz β, then its approximation by a constant function in an hypercube of length h gives a bias of order $d\,h^{\beta}$, and the constant can be estimated at the order $1/nh^d$, because there are nh^d observations in the interval. The latter improved rate (improved over $1/\sqrt{nh^d}$ as usual) leads to an optimal bandwidth of $h = \bar{h}(\beta) = n^{-\frac{1}{\beta+d}}$, and hence estimation rate $n^{-\frac{\beta}{\beta+d}}$, faster than usual. Typically, in any model, it is possible to improve the rate of convergence if the Fisher information is not finite.

In *regular statistical models*, the locally maximum likelihood approach is ideal (see Has'minskii and Ibragimov [1981] Chapter 1, Section 5 and recently Polzehl and Spokoiny [2006], Katkovnik and Spokoiny [2008]). But when the density of observations is discontinuous, an bayesian approach is often preferred (see Has'minskii and Ibragimov [1981], Chapter 1, Section 5). The study of local maximum likelihood estimator for model (4.1.1) stays an open problem.

Recently several approaches to the selection from the family of linear estimators were proposed, see for instance Goldenshluger and Lepski [2008], Goldenshluger and Lepski [2009a], Juditsky, Lepski, and Tsybakov [2009] and the references therein. These technologies are used in the construction of data-driven (adaptive) procedures and they are heavily based on the linearity property. As we already mentioned the locally bayesian estimators are non-linear and in Section 4.3 we propose the selection rule from this family. It requires, in particular, to develop new non-asymptotical exponential inequalities, which may have an independent interest.

Minimax estimation. The first part of the paper is devoted to the minimax over $\mathbb{H}_d(\beta, L, M, A)$ estimation. This means, in particular, that the parameters β, L, M and A are supposed to be known *a priori*. We find the *minimax rate of convergence* (4.1.4) on $\mathbb{H}_d(\beta, L, M, A)$ and propose the estimator being optimal in minimax sense (see Definition 14). Our first result in this direction consists in establishing a lower bound for maximal risk on $\mathbb{H}_d(\beta, L, M, A)$. We show that for any $\beta \in \mathbb{R}_+^*$, the minimax rate of convergence is bounded from below by the sequence

$$\varphi_n(\beta) = n^{-\frac{\beta}{\beta+d}}. \tag{4.1.4}$$

Next, we propose the minimax estimator, i.e. the estimator attaining the normalizing sequence (4.1.4). To construct the minimax estimator we use so-called *locally bayesian*

estimation construction which consists in the following. Let

$$V_h(y) = \bigotimes_{j=1}^{d} \big[y_j - h/2, y_j + h/2\big],$$

be the neighborhood around y such that $V_h(y) \subseteq [0,1]^d$, where $h \in (0,1)$ is a given scalar. Fix an integer number $b > 0$ and let

$$D_b = \sum_{m=0}^{b} \binom{m+d-1}{d-1}.$$

Let $\mathcal{P}_b = \{p = (p_1, \ldots, p_d) \,:\, p_i \in \mathbb{N},\ 0 \le |p| \le b\}$, $|p| = p_1 + \cdots + p_d$, we define the local polynomial

$$f_t(x) \;=\; \sum_{p \in \mathcal{P}_b} t_p \left(\frac{x-y}{h}\right)^p \mathbb{I}_{V_h(y)}(x), \quad x \in \mathbb{R}^d,\ t = (t_p \,:\, p \in \mathcal{P}_b), \tag{4.1.5}$$

where $z^p = z_1^{p_1} \cdots z_d^{p_d}$ for $z = (z_1, \ldots, z_d)$ and $\mathbb{I}$ denotes the indicator function. The local polynomial f_t can be viewed as an approximation of the regression function f inside of the neighborhood V_h and D_b the number of coefficients of this polynomial. Introduce the following subset of $\mathbb{R}^{D_b}$

$$\Theta\big(A, M\big) = \big\{t \in \mathbb{R}^{D_b} :\ 2t_{0,\ldots,0} - \|t\|_1 \ge A,\ \ \|t\|_1 \le M\big\}, \tag{4.1.6}$$

where $\|.\|_1$ is l_1-norm on $\mathbb{R}^{D_b}$. $\Theta(A, M)$ can be viewed as the set of coefficients t such that $A \le f_t(x) \le M$ for all $t \in \Theta(A, M)$ and for all x in the neighbourhood $V_h(y)$. Consider the *pseudo likelihood ratio*

$$L_h\big(t, Y^{(n)}\big) = \prod_{i:\, X_i \in V_h(y)} \big[f_t(X_i)\big]^{-1} \mathbb{I}_{\left[0, f_t(X_i)\right]}\big(Y_i\big), \quad t \in \Theta\big(A, M\big).$$

Set also

$$\pi_h(t) = \int_{\Theta(A,M)} \|t - u\|_1 \, L_h\big(u, Y^{(n)}\big) du, \quad t \in \Theta\big(A, M\big). \tag{4.1.7}$$

Let $\hat{\theta}(h)$ be the solution of the following minimization problem:

$$\hat{\theta}(h) = \arg \min_{t \in \Theta(A,M)} \pi_h(t). \tag{4.1.8}$$

The *locally bayesian estimator* $\bar{f}^h(y)$ of $f(y)$ is defined now as $\bar{f}^h(y) = \hat{\theta}_{0,\ldots,0}(h)$. Note that this local approach allows to estimate successive derivatives of function f. In this paper, only the estimation of f at a given point is studied.

We note that similar locally parametric approach based on maximum likelihood estimators was recently proposed in Polzehl and Spokoiny [2006] and Katkovnik and Spokoiny [2008] for *regular statistical models*.

As we see our construction contains an extra-parameter h to be chosen. To make this choice we use quite standard arguments. First, we note that in view of $f \in \mathbb{H}_d(\beta, L, M, A)$

$$\exists \theta = \theta(f, y, h) \in \Theta\big(A, M\big) \ : \quad \sup_{x \in V_h(y)} \big|f(x) - f_\theta(x)\big| \le L h^\beta .$$

For example, f_θ can be chosen like the Taylor polynomial (defined in (4.5.2)). With the property of isotropic Holder functions (see Definition 13), the last assertion is trivial. Thus, if h is chosen sufficiently small our original model (4.1.1) is well approximated inside of $V_h(y)$ by the "parametric" model

$$\mathcal{Y}_i = f_\theta(X_i) \times U_i, \quad i = 1, \dots, nh^d, \quad nh^d \in \mathbb{N}^*$$

in which the *bayesian estimator* $\hat{\theta}$ is rate-optimal (See Theorem 11).

It is worth mentioning that the analysis of the deviation of $(X_i, \mathcal{Y}_i)_{i=1,\dots nh^d}$ from $Y^{(nh^d)}$ is not simple. Namely here requirements $0 < A \le f(x) \le M, \forall x \in [0,1]^d$, are used. This assumption, which seems not to be necessary, allows us to make the presentation of basic ideas clear and to simplify routine computations (see also Remark 12). Without this assumption, the problem is open.

Finally, $h_n(\beta, L)$ is chosen as the solution of the following minimization problem

$$(nh^d)^{-1} + Lh^\beta \to \min_h \tag{4.1.9}$$

and we show that corresponding estimator $\bar{f}^{h_n(\beta,L)}(y)$ is minimax for $f(y)$ on $\mathbb{H}_d(\beta, L, M, A)$ if $\beta \le b$. Since the parameter $b > 0$ can be chosen in arbitrary way, the proposed estimator is minimax for any given value of the parameter $\beta > 0$.

Adaptive estimation. The second part of the paper is devoted to the adaptive minimax estimation over collection of isotropic functional classes in the model (4.1.1). At our knowledge, the problem of adaptive estimation in the multiplicative regression with the noise, having discontinuous density, is not studied in the literature.

Well-known drawback of minimax approach is the dependence of the minimax estimator on the parameters describing functional class on which the maximal risk is determined. In particular, the locally bayesian estimator $\bar{f}^h(\cdot)$ depends obviously on the parameters A and M via the solution of the minimization problem (4.1.8). Moreover $h_n(\beta, L)$ optimally chosen in view of (4.1.9) depends explicitly on β and L. To overcome this drawback the minimax adaptive approach was proposed (see Lepski [1990],Lepski [1991],Lepski, Mammen, and Spokoiny [1997]). The first question arising in the adaptation (reduced to the problem at hand) can be formulated as follows.

Does there exist an estimator which would be minimax on $\mathbb{H}(\beta, L, M, A)$ *simultaneously for all values of* β, L, A *and* M *belonging to some given subset of* $\mathbb{R}_+^4$ *?*

In section 4.3, we show that the answer to this question is negative, that is typical for the estimation of the function at a given point Lepski and Spokoiny [1997]. This answer can be reformulated in the following manner: the family of rates of convergence $\{\varphi_n(\beta),\ \beta \in \mathbb{R}_+^*\}$ is unattainable for the problem under consideration.

Thus, we need to find another family of normalizations for maximal risk which would be attainable and, moreover, optimal in view of some criterion of optimality. Nowadays, the most developed criterion of optimality is due to *Klutchnikoff* Klutchnikoff [2005].

We show that the family of normalizations, being optimal in view of this criterion, is

$$\phi_n(\beta) = \left(\frac{\rho_n(\beta)}{n}\right)^{\frac{\beta}{\beta+d}}, \qquad \rho_n(\beta) = 1 + \ln\left(\frac{\varphi_n(\beta)}{\varphi_n(b)}\right), \tag{4.1.10}$$

whenever $\beta \in]0, b]$. The factor ρ_n can be considered as *price to pay for adaptation* Lepski [1990].

The most important step in proving the optimality of the family (4.1.10) is to find an estimator, called *adaptive*, which attains the optimal family of normalizations. Obviously, we seek an estimator whose construction is *parameter-free*, i.e. independent of β, L, A and M. In order to explain our estimation procedure let us make several remarks.

First we note that the role of the constants A, M and β, L in the construction of the minimax estimator is quite different. Indeed, the constants A, M are used in order to determine the set $\Theta\big(A, M\big)$ needed for the construction of the locally bayesian estimator, see (4.1.7) and (4.1.8). However, this set does not depend on the localization parameter $h > 0$, in other words, the quantities A and M are not involved in the selection of optimal size of the local neighborhood given by (4.1.9). Contrary to that, the constants β, L are used for the derivation of the optimal size of the local neighborhood (4.1.9), but they are not involved in the construction of the collection of locally bayesian estimators $\big\{\hat{f}^h, h > 0\big\}$.

Next remark explains how to replace the unknown quantities A and M in the definition of $\Theta\big(A, M\big)$. Our first simple observation consists in the following: the estimator $\bar{f}^{h_n(\beta,L)}$ remains minimax if we replace $\Theta\big(A, M\big)$ in (4.1.7) and (4.1.8) by $\Theta\big(\tilde{A}, \tilde{M}\big)$ with any $0 < \tilde{A} \le A$ and $M \le \tilde{M} < \infty$. It follows from obvious inclusion $\mathbb{H}_d(\beta, L, A, M) \subseteq \mathbb{H}_d(\beta, L, \tilde{A}, \tilde{M})$. The next observation is less trivial and it follows from Proposition 4. Put $h_{\max} = n^{-\frac{1}{b+d}}$ and define for any function f

$$A(f) = \inf_{x \in V_{h_{\max}}(y)} f(x), \quad M(f) = \sum_{m=0}^{b} \sum_{p_1+\ldots+p_d=m} \left| \frac{\partial^m f(y)}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} \right|. \tag{4.1.11}$$

The following agreement will be used in the sequel: if the function f and $m \ge 1$ be such that $\partial^m f$ does not exist we will put formally $\partial^m f = 0$ in the definition of $M(f)$.

Then, $\bar{f}^{h_n(\beta,L)}$ remains minimax if we replace $\Theta\big(A, M\big)$ in (4.1.7) by $\Theta\big(A(f), M(f)\big)$. It remains to note that contrary to the quantities A and M the functionals $A(f)$ and $M(f)$ can be consistently estimated from the observation (4.1.1) and let $\hat{A}$ and $\hat{M}$ be the corresponding estimators. The idea now is to determine the collection of locally bayesian estimators $\big\{\hat{f}^h, h>0\big\}$ by replacing $\Theta\big(A, M\big)$ in (4.1.7) and (4.1.8) by the *random* parameter set $\hat{\Theta}$ which is defined as follows.

$$\hat{\Theta} = \Big\{ t \in \mathbb{R}^{D_b} : \ 2t_{0,\dots,0} - \|t\|_1 \geq 2^{-1}\hat{A}, \ \ \|t\|_1 \leq 4\hat{M} \Big\}.$$

In this context it is important to emphasize that the estimators $\hat{A}$ and $\hat{M}$ are built from the same observation which is used for the construction of the family $\big\{\hat{f}^h, h>0\big\}$.

Contrary to all saying above, the constants β and L cannot be estimated consistently. In order to select an "optimal" estimator from the family $\big\{\hat{f}^h, h>0\big\}$ we use general adaptation scheme due to *Lepski* Lepski [1990], Lepski [1992a]. To the best of our knowledge it is the first time when this method is applied in the statistical model with multiplicative noise and discontinuous distribution. Moreover, except already mentioned papers Polzehl and Spokoiny [2006] and Katkovnik and Spokoiny [2008], Lepski's procedure is typically applied to the selection from the collection of linear estimators (kernel estimators, locally polynomial estimator, etc.). In the present paper we apply this method to very complicated family of nonlinear estimators, obtained by the use of bayesian approach on the random parameter set. It required, in particular, to establish the exponential inequality for the deviation of locally bayesian estimator from the parameter to be estimated (Proposition 4). It generalizes the inequality proved for the parametric model Has'minskii and Ibragimov [1981] (Chapter 1, Section 5). This result seems to be new.

Simulations. The interest to the multiplicative models is explained by their popularity in the image processing, for example, SAR images Réfrégier [2002]. If the model is parametric then the use of the maximum likelihood estimator is rather standard at least if the likelihood is convex Bhattachar and Sundareshan [2000]. In the present paper we adopt the local parametric approximation to a purely non parametric model. As it proved, this strategy leads to the theoretically optimal statistical decisions. But the minimax as well as the minimax adaptive approach are asymptotical and it seems natural to check how proposed estimators work for reasonable sample size. In the simulation study, we test the bayesian estimator in the parametric and nonparametric cases. We show that the *adaptive* estimator approaches the *oracle* estimator. The *oracle* estimator is selected from the family $\Big\{\hat{f}^h,\ h>0\Big\}$ under the hypothesis f that is known. We show that the bayesian estimator performs well starting with $n \geq 100$.

This paper is organized as follows. In Section 4.2 we present the results concerning minimax estimation and Section 4.3 is devoted to the adaptive estimation. The simulations

are given in Section 4.4. The proofs of main results are proved in Section 4.5 (upper bounds) and section 4.6 (lower bounds). Auxiliary lemmas are postponed to Appendix (Section 4.7) contains the proofs of technical results.

4.2 Minimax estimation on isotropic Hölder class

In this section we present several results concerning minimax estimation. First, we establish lower bound for minimax risk defined on $\mathbb{H}_d(\beta, L, M, A)$ for any β, L, M and A. For any $(p_1, ..., p_d) \in \mathbb{N}^d$ we denote $p = (p_1, ..., p_d)$ and $|p| = p_1 + ... + p_d$.

Definition 13. *Fix $\beta > 0$, $L > 0$ and $M > 0$ and let $\lfloor \beta \rfloor$ be the largest integer strictly less than β. The* isotropic Hölder class $\mathbb{H}_d(\beta, L, M)$ *is the set of functions $f : [0,1]^d \to \mathbb{R}$ having on $[0,1]^d$ all partial derivatives of order $\lfloor \beta \rfloor$ and such that $\forall x, y \in [0,1]^d$*

$$\begin{aligned} \sum_{m=0}^{\lfloor \beta \rfloor} \sum_{|p|=m} \sup_{x \in [0,1]^d} \left| \frac{\partial^{|p|} f(x)}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} \right| &\leq M, \\ \left| \frac{\partial^{|p|} f(x)}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} - \frac{\partial^{|p|} f(y)}{\partial y_1^{p_1} \cdots \partial y_d^{p_d}} \right| &\leq L \big[\|x - y\|_1 \big]^{\beta - \lfloor \beta \rfloor}, \quad \forall\, |p| = \lfloor \beta \rfloor. \end{aligned}$$

This definition implies that if $f \in \mathbb{H}_d(\beta, L, M, A)$, then $A \leq A(f)$ and $M(f) \leq M$, where $\mathbb{H}_d(\beta, L, M, A)$ is defined in the beginning of this paper. $A(f)$ and $M(f)$ are defined in (4.1.11).

Maximal and minimax risk on $\mathbb{H}_d(\beta, L, M, A)$. To measure the performance of estimation procedures on $\mathbb{H}_d(\beta, L, M, A)$ we will use minimax approach.

Let $\mathbb{E}_f = \mathbb{E}_f^n$ be the mathematical expectation with respect to the probability law of the observation $Y^{(n)}$ satisfying (4.1.1). We define first the maximal risk on $\mathbb{H}_d(\beta, L, M, A)$ corresponding to the estimation of the function f at a given point $y \in [0,1]^d$.
Let $\tilde{f}$ be an arbitrary estimator built from the observation $Y^{(n)}$. Let $\forall q > 0$

$$R_{n,q}\big[\tilde{f}, \mathbb{H}_d(\beta, L, M, A)\big] = \sup_{f \in \mathbb{H}_d(\beta, L, M, A)} \mathbb{E}_f \big| \tilde{f}(y) - f(y) \big|^q.$$

The quantity $R_{n,q}\big[\tilde{f}, \mathbb{H}_d(\beta, L, M, A)\big]$ is called *maximal risk* of the estimator $\tilde{f}$ on $\mathbb{H}_d(\beta, L, M, A)$ and the *minimax risk* on $\mathbb{H}_d(\beta, L, M, A)$ is defined as

$$R_{n,q}\big[\mathbb{H}_d(\beta, L, M, A)\big] = \inf_{\tilde{f}} R_{n,q}\big[\tilde{f}, \mathbb{H}_d(\beta, L, M, A)\big],$$

where inf is taken over the set of all estimators.

Definition 14. *The normalizing sequence ψ_n is called minimax rate of convergence (MRT) and the estimator $\hat{f}$ is called minimax (asymptotically minimax) if*

$$\begin{aligned}\liminf_{n\to\infty} \psi_n^{-q}\, R_{n,q}\big[\hat{f}, \mathbb{H}_d(\beta, L, M, A)\big] &> 0;\\ \limsup_{n\to\infty} \psi_n^{-q}\, R_{n,q}\big[\hat{f}, \mathbb{H}_d(\beta, L, M, A)\big] &< \infty.\end{aligned}$$

Theorem 10. *For any $\beta > 0$, $L > 0$, $M > 0$, $A > 0$, $q \geq 1$ and $d \geq 1$*

$$\liminf_{n\to\infty} \varphi_n^{-q}(\beta) R_{n,q}\big[\mathbb{H}_d(\beta, L, M, A)\big] > 0, \qquad \varphi_n(\beta) = n^{-\frac{\beta}{\beta+d}}.$$

Remark 12. *The obtained result shows that on $\mathbb{H}_d\big(\beta, L, M, A\big)$ the minimax rate of convergence cannot be faster than $n^{-\frac{\beta}{\beta+d}}$. In view of the obvious inclusion $\mathbb{H}_d(\beta, L, M, A) \subset \mathbb{H}_d(\beta, L, M)$ the minimax rate of convergence on an isotropic Hölder class is also bounded from below by $n^{-\frac{\beta}{\beta+d}}$.*

The next theorem shows how to construct the minimax estimator basing on locally bayesian approach. Put $\bar{h} = (Ln)^{-\frac{1}{\beta+d}}$ and let $\bar{f}^{\bar{h}}(y) = \hat{\theta}_{0,\dots,0}\big(\bar{h}\big)$ is given by (4.1.6), (4.1.7) and (4.1.8) with $h = \bar{h}$.

Theorem 11. *Let $\beta > 0$, $L > 0$, $M > 0$ and $A > 0$ be fixed. Then there exists the constant C_* such that for any $n \in \mathbb{N}^*$ satisfying $n\bar{h}^d \geq \big(\lfloor\beta\rfloor + 1\big)^d$*

$$\varphi_n^{-q}(\beta) R_{n,q}\Big[\bar{f}^{\bar{h}}(y), \mathbb{H}_d(\beta, L, M, A)\Big] \leq C^*, \quad \forall q \geq 1.$$

The explicit form of C^* is given in the proof.

Remark 13. *We deduce from Theorems 10 and 11 that the estimator $\bar{f}^{\bar{h}}(y)$ is minimax on $\mathbb{H}_d(\beta, L, M, A)$.*

4.3 Adaptive estimation on isotropic Hölder classes

This section is devoted to the adaptive estimation over the collection of the classes $\Big\{\mathbb{H}_d(\beta, L, M, A)\Big\}_{\beta,L,M,A}$. We will not impose any restriction on possible values of L, M, A, but we will assume that $\beta \in (0, b]$, where b, as previously, is an arbitrary a priori chosen integer.

We start with formulating the result showing that there is no optimally adaptive estimator (here we follow the terminology introduced in Lepski [1991], Lepski [1992a]). It means that there is no an estimator which would be minimax simultaneously for several values of parameter β even if all other parameters L, M and A are supposed to be fixed. This result does not require any restriction on β as well.

Theorem 12. *For any subset of* $\mathbb{R}^+ \setminus \{0\}$ *noted* $\mathbb{B}$ *such that card*$(\mathbb{B}) \geq 2$, *for any* $\beta_1, \beta_2 \in \mathbb{B}$ *and any* $L > 0,\ M > 0,\ A > 0$

$$\liminf_{n\to\infty} \inf_{\tilde{f}} \Big[\varphi_n^{-q}(\beta_1)\, R_{n,q}\big(\tilde{f}, \mathbb{H}_d(\beta_1, L, M, A)\big) + \varphi_n^{-q}(\beta_2)\, R_{n,q}\big(\tilde{f}, \mathbb{H}_d(\beta_2, L, M, A)\big)\Big] = +\infty,$$

where inf *is taken over all possible estimators.*

The assertion of Theorem 12 can be considerably specified if $\mathbb{B} = (0, b]$. To do that we will need the following definition. Let $\Psi = \{\psi_n(\beta)\}_{\beta\in(0,b]}$ be a given family of normalizations.

Definition 15. *The family* Ψ *is called admissible if there exist an estimator* $\hat{f}_n$ *such that for* some $L > 0, M > 0$ *and* $A > 0$

$$\limsup_{n\to\infty} \psi_n^{-q}(\beta)\, R_{n,q}\big(\hat{f}, \mathbb{H}_d(\beta, L, M, A)\big) < \infty, \quad \forall \beta \in (0, b]. \tag{4.3.1}$$

The estimator $\hat{f}_n$ *satisfying (4.3.1) is called* Ψ-attainable. *The estimator* $\hat{f}_n$ *is called* Ψ-adaptive *if (4.3.1) holds for* any $L > 0, M > 0$ *and* $A > 0$.

Note that the result proved in Theorem 12 means that the family of rates of convergence $\{\varphi_n(\beta)\}_{\beta\in(0,b]}$ is not admissible. Let Φ be the following family of normalizations:

$$\phi_n(\beta) = \left(\frac{\rho_n(\beta)}{n}\right)^{\frac{\beta}{\beta+d}}, \quad \rho_n(\beta) = 1 + \ln\left(\frac{\varphi_n(\beta)}{\varphi_n(b)}\right), \ \beta \in (0, b].$$

We remark that $\phi_n(b) = \varphi_n(b)$ and $\rho_n(\beta) \sim \ln n$ for any $\beta \neq b$.

Theorem 13. *Let* $\Psi = \{\psi_n(\beta)\}_{\beta\in(0,b]}$ *be an arbitrary admissible family of normalizations.*
I. *For any* $\alpha \in (0, b]$ *such that* $\psi_n(\alpha) \neq \varphi_n(\alpha)$, *there exists an admissible family* $\{\upsilon_n(\beta)\}_{\beta\in(0,b]}$ *for which*

$$\lim_{n\to\infty} \upsilon_n(\alpha)\psi_n^{-1}(\alpha) = 0.$$

II. *If there exists* $\gamma \in (0, b)$ *such that*

$$\lim_{n\to\infty} \psi_n(\gamma)\phi_n^{-1}(\gamma) = 0, \tag{4.3.2}$$

then necessarily

$$\textbf{(a)} \qquad \lim_{n\to\infty} \psi_n(\beta)\phi_n^{-1}(\beta) > 0, \quad \forall \beta \in (0, \gamma);$$

$$\textbf{(b)} \qquad \lim_{n\to\infty} \left[\frac{\psi_n(\gamma)}{\phi_n(\gamma)}\right]\left[\frac{\phi_n(\beta)}{\psi_n(\beta)}\right] = 0, \quad \forall \beta \in (\gamma, b].$$

Several remarks are in order.

1. We note that if the family of normalizations Φ is admissible, i.e. one can construct Φ-attainable estimator, then Φ is in an *optimal* family of normalizations in view of Kluchnikoff criterion Klutchnikoff [2005]. It follows from the second assertion of the theorem. We note however that a Φ-attainable estimator may depend on $L > 0, M > 0$ and $A > 0$, and, therefore, this estimator have only theoretical interest. In the next section we construct Φ-adaptive estimator, which is, by its definition, fully parameter-free. Moreover, this estimator obviously proves that Φ is admissible, and, therefore, optimal as it was mentioned above.

2. The assertions of Theorem 13 allows us to give rather simple interpretation of Kluchnikoff criterion. Indeed, the first assertion, which is easily deduced from Theorem 12, shows that any admissible family of normalizations can be improved by another admissible family at any given point $\alpha \in (0, b]$ except maybe one. In particular, it concerns the family Φ if it is admissible. On the other hand, the second assertion of the theorem shows that there is no admissible family which would outperform the family Φ at two points. Moreover, in view of (**b**), Φ-adaptive (attainable) estimator, if exists, has the same precision on $\mathbb{H}_d(\beta, L, M, A)$, $\beta < \gamma$, as any Ψ-adaptive(attainable) estimator whenever Ψ satisfies (4.3.2). Additionally, (**a**) implies that the gain in the precision provided by Ψ-adaptive (attainable) estimator on $\mathbb{H}_d(\gamma, L, M, A)$ leads automatically to much more losses on $\mathbb{H}_d(\beta, L, M, A)$ for any $\beta > \gamma$ with respect to the precision provided by Φ-adaptive(attainable) estimator. We conclude that Φ-adaptive(attainable) estimator outperforms any Ψ-adaptive(attainable) estimator whenever Ψ satisfies (4.3.2). It remains to note that any admissible family not satisfying (4.3.2) is asymptotically equivalent to Φ.

Construction of Φ-adaptive estimator. As it was already mentioned in Introduction the construction of our estimation procedure consists of several steps. First, we determine the set $\hat{\Theta}$, built from observation, which is used after that in order to define the family of locally bayesian estimators. Next, based on Lepski's method (see Lepski [1991] and Lepski, Mammen, and Spokoiny [1997]), we propose data-driven selection from this family.

First step: Determination of parameter set. Put $h_{\max} = n^{-\frac{1}{b+d}}$ and let $\tilde{\theta}$ be the solution of the following minimization problem.

$$\inf_{t \in \mathbb{R}^{D_b}} \sum_{i: X_i \in V_{\max}(y)} \left[2Y_i - t\, K^\top \left(\frac{X_i - y}{h_{\max}} \right) \right]^2, \quad V_{\max}(y) = V_{h_{\max}}(y), \tag{4.3.3}$$

where the D_b-dimensional vector $K(z) = (z^p \ : \ p \in \mathcal{P}_b)$ and the sign $\top$ below means the transposition. Thus, $\tilde{\theta}$ is the local least squared estimator and its explicit expression is given

by

$$\tilde{\theta} = 2\left[\sum_{i:X_i\in V_{\max}(y)} K^{\top}\left(\frac{X_i-y}{h_{\max}}\right)K\left(\frac{X_i-y}{h_{\max}}\right)\right]^{-1}\big[\mathcal{K}_n(y)\big]^{\top}Y, \tag{4.3.4}$$

where $Y=(Y_1,\dots,Y_n)$ and $\mathcal{K}_n(y)=\Big[K^{\top}\left(\frac{X_i-y}{h_{\max}}\right)\mathbb{I}_{V_{\max}(y)}(X_i)\Big]_{i=1,\dots n}$ is the design matrix. Put

$$\tilde{\delta}_p = p_1!...p_d!\, h_{\max}^{-|p|}\,\tilde{\theta}_p, \quad |p|\le b.$$

Introduce the following quantities

$$\hat{A}=\tilde{\delta}_{0\dots,0}, \qquad \hat{M}=\big\|\tilde{\delta}\big\|_1, \tag{4.3.5}$$

and define the random parameter set as follows.

$$\hat{\Theta}=\Big\{t\in\mathbb{R}^{D_b}:\ 2t_{0,\dots,0}-\|t\|_1\ge 2^{-1}\hat{A},\ \|t\|_1\le 4\hat{M}\Big\}. \tag{4.3.6}$$

Second step: Collection of locally bayesian estimators. Put

$$\hat{\pi}_h(t) \ = \ \int_{\hat{\Theta}}\|t-u\|_1\,L_h\big(u,Y^{(n)}\big)du; \tag{4.3.7}$$

$$\hat{\theta}^*(h) \ = \ \arg\min_{t\in\hat{\Theta}}\hat{\pi}_h(t). \tag{4.3.8}$$

The family of locally bayesian estimator $\hat{\mathcal{F}}$ is defined now as follows.

$$\hat{\mathcal{F}}=\Big\{\hat{f}^h(y)=\hat{\theta}^*_{0,\dots,0}(h),\ h\in\big(0,h_{\max}\big]\Big\}. \tag{4.3.9}$$

Third step: Data-driven selection from the collection $\hat{\mathcal{F}}$. Put

$$h_k=2^{-k}h_{\max},\ \ k=0,\dots,\mathrm{k}_n,$$

where k_n is smallest integer such that $h_{\mathrm{k}_n}\ge h_{\min}=\ln^{\frac{b}{d(b+d)}}n^{-1/d}$. Set

$$\hat{\mathcal{F}}^*=\Big\{\hat{f}^{(k)}(y)=\hat{\theta}^*_{0,\dots,0}(h_k),\ \ k=0,\dots,\mathrm{k}_n\Big\}.$$

We put $\hat{f}^*(y)=\hat{f}^{(\hat{k})}(y)$, where $\hat{f}^{(\hat{k})}(y)$ is selected from $\hat{\mathcal{F}}^*$ in accordance with the rule:

$$\hat{k}=\inf\Big\{k=\overline{0,\mathrm{k}_n}:\ \big|\hat{f}^{(k)}(y)-\hat{f}^{(l)}(y)\big|\le\hat{M}S_n\big(l\big),\ \ l=\overline{k+1,\mathrm{k}_n}\Big\}. \tag{4.3.10}$$

Here we have used the following notations.

$$S_n(l)=432D_b^3(32qd+16)\,\lambda_n^{-1}\big(h_l\big)\,\left[\frac{1+l\ln 2}{n\big(h_l\big)^d}\right], \quad l=0,1,\dots,\mathrm{k}_n, \tag{4.3.11}$$

and $\lambda_n(h)$ is the smallest eigenvalue of the matrix

$$\mathcal{M}_{nh}(y) = \frac{1}{nh^d}\sum_{i=1}^{n} K^\top\left(\frac{X_i - y}{h}\right) K\left(\frac{X_i - y}{h}\right) \mathbb{I}_{V_h(y)}(X_i), \tag{4.3.12}$$

which is completely determined by the design points and by the number of observations. We will prove that there exists a nonnegative real λ, such that $\lambda_n(h) \geq \lambda$ for any $n \geq 1$ and any $h \in \left[h_{\min}, h_{\max}\right]$ (see Lemma 9).

Theorem 14. *Let an integer number $b > 0$ be fixed. Then for any $\beta \in (0, b]$, $L > 0$, $M > 0, A > 0$ and $q \geq 1$*

$$\limsup_{n\to\infty} \phi_n^{-q}(\beta) R_{n,q}\left[\hat{f}^*(y), \mathbb{H}_d(\beta, L, M, A)\right] < \infty.$$

Remark 14. *The assertion of the theorem means that the proposed estimator $\hat{f}^*(y)$ is Φ-adaptive. It implies in particular that the family of normalizations Φ is admissible. This, together with Theorem 13 allows us to state the optimality of Φ in view of Kluchnikoff criterion (see Klutchnikoff [2005]).*

4.4 Simulation study

We will consider the case $d = 1$. The data are simulated accordingly to the model (4.1.1), where we use the following functions (Figure 4.1).

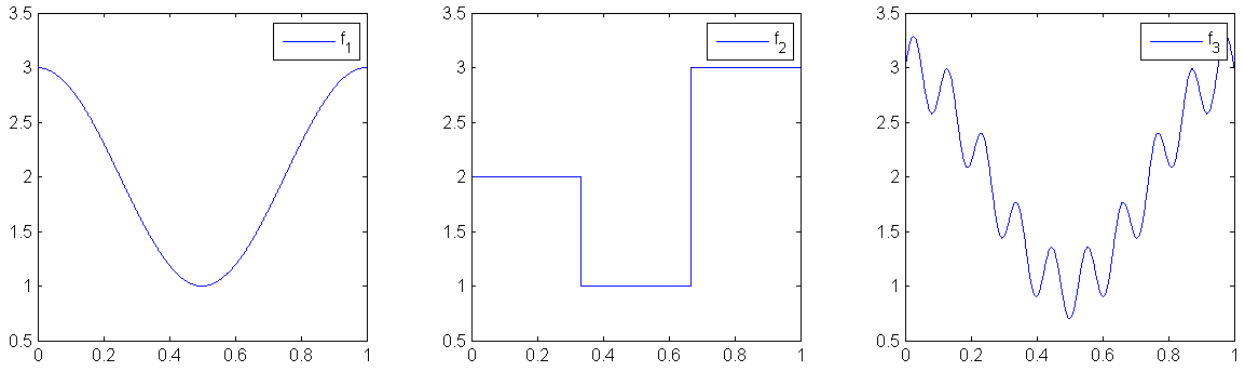


Figure 4.1:

Here $f_1(x) = \cos(2\pi x) + 2$, $f_2(x) = 2.\mathbb{I}_{[x\leq 1/3]} + 1.\mathbb{I}_{[1/3<x\leq 2/3]} + 3.\mathbb{I}_{[2/3<x]}$ and $f_3(x) = \cos(2\pi x) + 2 + 0.3\sin(19\pi x)$

To construct the family of estimators we use the linear approximation ($b = 2$), i.e. within the neighbourhoods of the given size h, the locally bayesian estimator has the form

$$\hat{f}^h(x) = \hat{\theta}_0 + \hat{\theta}_1 x, \quad x \in [0, 1].$$

We define the ideal (oracle) value of the parameter $\tilde{h} = \tilde{h}(f)$ as the minimizer of the risk:

$$\tilde{h} = \arg \inf_{h \in [1/n,1]} \mathbb{E}_f \big| \hat{f}^h(y) - f(y) \big|.$$

To compute it we apply Monte-Carlo simulations (10000 repetitions). Our first objective is to compare the risk provided by the "oracle" estimator $\hat{f}^{\tilde{h}}(\cdot)$ and whose provided by the adaptive estimator from Section 4.3. Figure 4.2 shows the deviation of the adaptive estimator from the function to be estimated. In several points, for example in $y = 1/2$, we remark so-called over-smoothing phenomenon, inherent to any adaptive estimator.

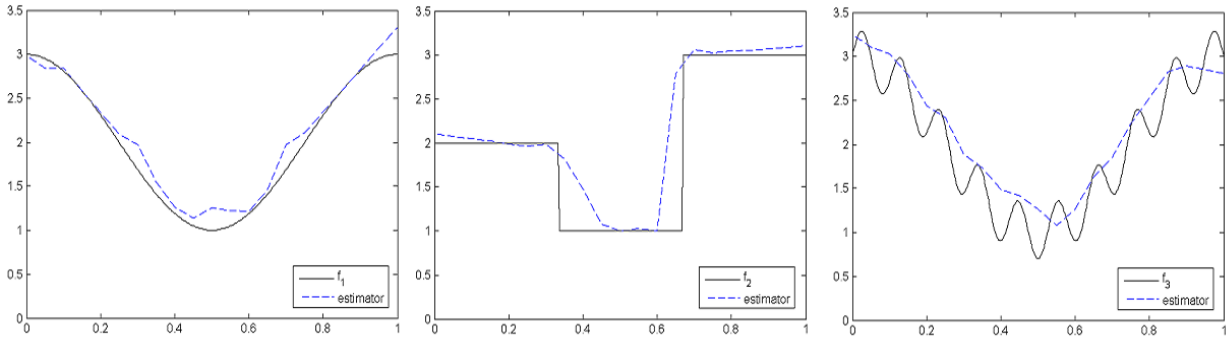


Figure 4.2:

Oracle-adaptive ratio. We compute the risks of the oracle and the adaptive estimator in 100 points of the interval $(0, 1)$. The next tabular presents the mean value of the ratio oracle risk/adaptive risk calculated for the functions f_1, f_2, f_3 and $n = 100, 1000$.

	n = 100		n = 1000	
function	adaptive risk	oracle-adaptive ratio	adaptive risk	oracle-adaptive ratio
f_1	0.13	0.84	0.03	0.85
f_2	0.3	0.71	0.1	0.75
f_3	0.28	0.65	0.2	0.68

Figure 4.3 presents the "oracle risk/adaptive risk" ratio as the function of the number of observations n.

Adaptation versus parametric estimation. We consider the function f_4 (figure 4.4), which is linear inside the neighborhood of size $h_* = 1/8$ around point $1/2$ and simulate $n = 1000$ observations in accordance with the model (4.1.1). Using only the observations corresponding to the interval $[3/8, 5/8]$ we construct the bayesian estimator $\hat{f}^{1/8}(1/2)$.

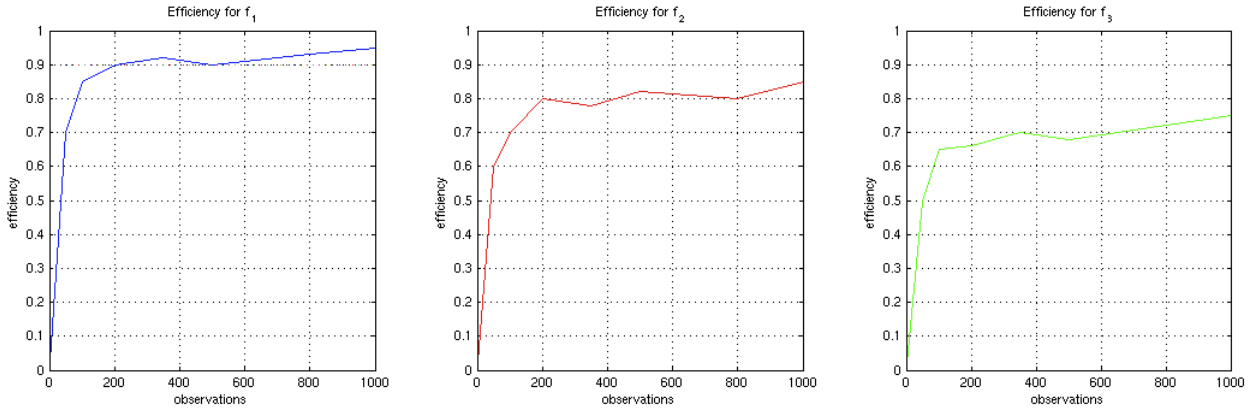


Figure 4.3:

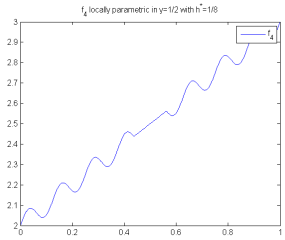


Figure 4.4:

It is important to emphasize that this estimator is efficient Has'minskii and Ibragimov [1981] since the model is parametric. Our objective now is to compare the risk of our adaptive estimator with the risk provided by the estimator $\hat{f}^{1/8}(1/2)$. We also try to understand how far is the localization parameter $h_{\hat{k}}$, inherent to the construction of our adaptive estimator, from the true value $1/8$. We compute the risk of each estimator via Monte-Carlo method with 10000 repetitions. For each repetition the procedure select the adaptive bandwidth $h_{\hat{k}}^{(j)}$, $j = 1, ..., 10000$. We confirm once again the over-smoothing phenomenon since

$$h_{\hat{k}}^{(j)} \sim 0.1405 > h_* = 0.1250, \quad j = 1, ..., 10000.$$

Note however that the adaptive procedure selects the neighborhood of the size which is quite close to the true one. We also compute the risks of both estimators: "bayesian risk"=**0.0206** and "adaptive risk=**0.0308**". We conclude that the estimation accuracy provided by our adaptive procedure is quite satisfactory.

4.5 Proofs of main results: upper bounds

Let $\mathcal{H}_n, n > 1$ be the following subinterval of $(0, 1)$.

$$\mathcal{H}_n = \left[\frac{(b+1) \vee \left(\ln n\right)^{\frac{1}{(d+d^2)}}}{n^{1/d}}, \left(\frac{1}{\ln n}\right)^{\frac{1}{b+d}} \right]. \tag{4.5.1}$$

Later on we will consider only the values of h belonging to $\mathcal{H}_n$. We begin by stating the control of large deviations which is necessary to prove minimax and adaptive results.

4.5.1 Auxiliary results

Introduce the following notations. For any $h > 0$ satisfying (4.5.1) put $\omega = \omega(f, y, h) = \left\{\omega_p : \ p \in \mathcal{P}_b\right\}$, where $\omega_0 = \omega_{0,\dots,0} = f(y)$ and

$$\omega_p = \frac{\partial^{|p|} f(y)}{\partial y_1^{p_1} \cdots \partial y_d^{p_d}} \frac{h^{|p|}}{p_1! \dots p_d!}, \ p\mathcal{P}_b. \tag{4.5.2}$$

Remind the agreement which we follow in the present paper: if the function f and vector p are such that $\partial^{|p|} f$ does not exist we put $\omega_p = 0$.

Let $f_\omega(x)$, given by (4.1.5), be the local polynomial approximation of f inside $V_h(y)$ and let b_h be the corresponding approximation error, i.e.

$$b_h = \sup_{x \in V_h(y)} \left| f_\omega(x) - f(x) \right|. \tag{4.5.3}$$

It is easy to show that $b_h \le Ldh^\beta$ by definition of ω in (4.5.2). Put finally the bias-variance ratio and an other term which are necessary for the interpretation of the following results.

$$\mathcal{N}_h = b_h \times nh^d, \quad \mathcal{E}(h) = \exp\left\{ \frac{(1 + 6D_b^2)\mathcal{N}_h}{6A(f)D_b^2} \right\}. \tag{4.5.4}$$

Introduce the random events $G_{\hat{M}} = \left\{ \left| \hat{M} - M(f) \right| \le M(f)/2 \right\}$ and $G_{\hat{A}} = \left\{ \left| \hat{A} - A(f) \right| \le A(f)/2 \right\}$ and put $G = G_{\hat{M}} \cap G_{\hat{A}}$ where $\hat{A}$ and $\hat{M}$ are defined in (4.3.5) in Section 4.3. Recall that $\lambda_n(h)$ (defined in Section 4.3) is the smallest eigenvalue of the matrix

$$\mathcal{M}_{nh}(y) = \frac{1}{nh^d} \sum_{i=1}^n K^\top \left(\frac{X_i - y}{h} \right) K \left(\frac{X_i - y}{h} \right) \mathbb{I}_{V_h(y)}(X_i),$$

and $K(z)$ is the D_b-dimensional vector of the monomials z^p, $p \in \mathcal{P}_b$.

The next proposition is the milestone for all results proved in the paper.

Proposition 4. *For any h satisfying (4.5.1) and any f such that $A(f) > 0$ and $M(f) < \infty$, then $\forall \varepsilon > 0$*

$$\mathbb{P}_f\left(nh^d\big|\hat{f}^h(y) - f(y)\big| \geq \varepsilon,\, G\right) \leq \mathfrak{B}\big(A(f), M(f)\big)\mathcal{E}(h)\exp\left\{-\frac{\lambda_n(h)\,\varepsilon}{432M(f)\,D_b^3}\right\},$$

where $\hat{f}^h(y) \in \hat{\mathcal{F}}$ defined in (4.3.9). The explicit expression of the function $\mathfrak{B}(\cdot,\cdot)$ is given in the beginning of the proof of the proposition.

In seeking this type of result (control of large deviations), the new point is to control the term $\mathcal{E}(h)$ which can be very large compared to the normalization $(nh^d)^{-1}$. It is a consequence that the bayesian estimator is non-linear. The bias-variance ratio $\mathcal{N}_h$ (defined in (4.5.4)), which appears in $\mathcal{E}(h)$, allows to obtain similar results with linear estimators (including the choice of the bandwidth).

The next proposition provides us with upper bound for the risk of a locally bayesian estimator.

Proposition 5. *For any $n \in \mathbb{N}^*$, $h \in \mathcal{H}_n$ and any $f \in \mathbb{H}_d(\beta, L, M, A)$, then $\exists \lambda > 0$ does not depend of n such that:*

$$\mathbb{E}_f\big|\hat{f}^h(y) - f(y)\big|^q \mathbb{I}_G \leq C_q^*\left[\frac{1 \vee Ld\, nh^{\beta+d}}{nh^d}\right]^q, \quad q \geq 1.$$

where

$$C_q^* = \frac{\mathfrak{B}(A,M)}{q}\left(\frac{144M(f)\,D_b}{A(f)\lambda(1+6D_b^2)^{-1}}\right)^q \int_0^{+\infty}(\eta+1)^{q-1}\exp\left\{-\frac{\lambda\,\eta}{432M(f)\,D_b^3}\right\}.$$

4.5.2 Proof of Proposition 4

Before beginning this proof, we introduce several notations which simplify understanding and we state Lemma 8. The knowledge of the likelihood L_h is not directly involved in the proof of Proposition 1, but only in the proof of Lemma 8. In fact, the proof of Proposition 4 requires only Lemma 8 and using the bayesian estimator with the local non-parametric approach. For it, we use standard arguments like Markov inequality or the following inclusion

$$\left\{nh^d\big\|\hat{\theta}^*(h) - \theta\big\|_1 \geq \varepsilon\right\} \subseteq \left\{\inf_{nh^d\|t-\theta\|_1 \geq \varepsilon} \hat{\pi}_h(t) \leq \hat{\pi}_h(\theta)\right\}.$$

where $\hat{\theta}^*(h)$ minimizes $\hat{\pi}_h$ defined in (4.3.7).

Set for $z > 0, u > 0$ and $v > 0$ Define finally for any $u > 0, v > 0$

$$\mathfrak{B}(u,v) = \sup_{z \geq 0} 16e\big(1 \vee D_b u^{-1}\big)\Sigma^*(v)\,[\mathcal{B}_z + 6]\exp\left\{-\frac{\lambda z}{432 v D_b^3}\right\}, \tag{4.5.5}$$

where $\mathcal{B}_z = z^{D_b+1} + 2^{D_b}(2z+2)^{\frac{D_b}{2}} + 6 + D_b\left(z^{D_b} + (2z+2)^{\frac{D_b}{2}-1}\right)$, $\lambda > 0$ is described in Lemma 9 and does not depend of n, and

$$\Sigma^*(v) = \frac{c^2(v)\big(3-c(v)\big)}{\big(1-c(v)\big)^3}, \quad c(v) = \exp\big\{-(54vD_b^2)^{-1}\big\}. \tag{4.5.6}$$

Auxiliary Lemma. Introduce the vector $\theta = \theta(f,y,h) = \big\{\theta_p : p \in \mathcal{P}_b\big\}$, where

$$\theta_0 = \theta_{0,\dots,0} = \omega_0 + b_h, \quad \theta_p = \omega_p, \ \ |p| \neq 0, \tag{4.5.7}$$

where ω is the coefficients of Taylor polynomial defined in (4.5.2). The definition of b_h implies obviously

$$f_\theta(x) \ \geq \ f(x), \quad \forall x \in V_h(y). \tag{4.5.8}$$

This trivial inequality will be extensively exploited in the sequel.

Let $U_n = \big\{u \in \mathbb{R}^{D_b} \ : \ u = nh^d(t-\theta), \ t \in \Theta(A(f)/4, 9M(f))\big\}$. For any $u \in U_n$, put the process

$$Z_{h,\theta}(u) = \frac{L_h\big(\theta + (nh^d)^{-1}u, Y^{(n)}\big)}{L_h\big(\theta, Y^{(n)}\big)}.$$

Note that in view of (4.5.8), the event $Y_i \leq f_\theta(X_i)$ is always realized, because the multiplicative noise belongs to $[0,1]$, therefore $Y_i \leq f(X_i) \leq f_\theta(X_i)$. By definition of L_h, the process $Z_{h,\theta}$ can be written as follows

$$Z_{h,\theta}(u) = \prod_{i:\, X_i \in V_h(y)} \frac{f_\theta(X_i)}{f_{\theta+u(nh^d)^{-1}}(X_i)}\, \mathbb{I}_{\big[Y_i \leq f_{\theta+u(nh^d)^{-1}}(X_i)\big]}, \quad u \in U_n. \tag{4.5.9}$$

Lemma 8. *For any $f \in \mathbb{H}_d(\beta, L, M, A)$ and $h \in \mathcal{H}_n$*

1. $\sup\limits_{u_1,u_2 \in U_n} \|u_1 - u_2\|_1^{-1} \mathbb{E}_f \big|Z_{h,\theta}(u_1) - Z_{h,\theta}(u_2)\big| \leq \mathcal{C}_h$,
2. $\mathbb{E}_f Z_{n,\theta}^{1/2}(u) \leq e^{-g_h\left(\|u\|_1\right)}, \ \ \forall u \in U_n$,
3. $\mathbb{P}_f\left\{\int_{[0,\delta]^{D_b}} Z_{h,\theta}(u)du < \frac{\delta^{D_b}}{2}\right\} < 2\mathcal{C}_h\delta, \ \ \forall \delta > 0.$

where

$$\mathcal{C}_h = 8\big(1 \vee D_b A^{-1}(f)\big)\exp\big\{1 + \mathcal{N}_h/A(f)\big\}, \quad g_h(a) = \frac{\lambda_n(h)a}{18M(f)D_b} - \frac{\mathcal{N}_h}{A(f)}.$$

$\lambda_n(h)$ is the smallest eigenvalue of the matrix $\mathcal{M}_{nh}(y)$ defined in (4.3.12).

The knowledge of the likelihood occurs only in the proof of previous lemma.

Proof of Proposition 4. This proof is based on a result, already known in nonparametric estimation, given by Has'minskii and Ibragimov [1981] in their book (Chapter 1, Section 5, Theorem 5.2). We begin by stating two assertions which will prove in the first and second parts of this proof. Finally, we conclude easily with these two assertions and the Markov inequality. Let us introduce a new process z_h which is the renormalized process $Z_{h,\theta}$.

$$z_h(u) = \frac{Z_{h,\theta}(u)}{\int_{\hat{U}_n} Z_{h,\theta}(v)dv}, \quad u \in \hat{U}_n$$

where $\hat{U}_n = nh^d\big(\hat{\Theta} - \theta\big)$ where the set $\hat{\Theta}$ is defined in (4.3.6). Remind that $\theta = \theta(f, y, h)$ is defined in (4.5.7).

Assertion 1. *For any $\varepsilon > 0$, and for all r such that $0 < r < \varepsilon/3$, we assume*

$$\mathbb{P}_f\left(nh^d\big|\hat{f}^h(y) - f(y)\big| \geq \varepsilon, G\right) \leq 2\mathbb{P}_f\left(\int_{\hat{U}_n(r)} \|u\|_1 z_h(u)du > \frac{r}{2}, G\right).$$

The proof of Assertion 1 use the definition of the process z_h and the following inclusion

$$\left\{nh^d\big\|\hat{\theta}^*(h) - \theta\big\|_1 \geq \varepsilon\right\} \subseteq \left\{\inf_{nh^d\|t-\theta\|_1 \geq \varepsilon} \hat{\pi}_h(t) \leq \hat{\pi}_h(\theta)\right\}.$$

where $\hat{\theta}^*(h)$ minimizes $\hat{\pi}_h$ defined in (4.3.7).

Assertion 2. *For all $h \in \mathcal{H}_n$ and any f such that $A(f) > 0$ and $M(f) < \infty$, then for any $a > 0$*

$$\mathbb{E}_f\left[\int_{\hat{U}_n \cap \left\{\|u\|_1 > a\right\}} \|u\|_1 z_h(u)\, du\, \mathbb{I}_G\right] \leq a\Sigma^*\big(M(f)\big)\, \mathcal{B}_a\, \mathcal{C}_h \exp\left\{-\frac{1}{6D_b} g_h(a)\right\},$$

where $g_h(.)$ is defined in Lemma 8.

This assertion requires Markov inequality and Lemma 8.

First step: Proof of Assertion 1.
The definition of $\hat{\theta}^*(h)$ and $\theta = \theta(f, y, h)$ implies $\forall \varepsilon > 0$

$$\begin{aligned} \mathbb{P}_f\left(nh^d\big|\hat{f}^h(y) - f(y)\big| \geq \varepsilon, G\right) &\leq \mathbb{P}_f\left(nh^d\big|\hat{\theta}_0^*(h) - \theta_0\big| \geq \varepsilon, G\right) \\ &\leq \mathbb{P}_f\left(nh^d\big\|\hat{\theta}^*(h) - \theta\big\|_1 \geq \varepsilon, G\right). \end{aligned} \tag{4.5.10}$$

Some remarks are in order. First, it is easily seen that $\theta \in \Theta\big(A(f), 3M(f)\big)$. Therefore, if the event G holds then $\theta \in \hat{\Theta}$. Remind also that $\hat{\theta}^*(h)$ minimizes $\hat{\pi}_h$ defined in (4.3.7) and, therefore, the following inclusion holds since $\hat{\theta}^*(h) \in \hat{\Theta}$.

$$\left\{\left(nh^d\big\|\hat{\theta}^*(h) - \theta\big\|_1 \geq \varepsilon\right) \cap G\right\} \subseteq \left\{\left(\inf_{nh^d\|t-\theta\|_1 \geq \varepsilon} \hat{\pi}_h(t) \leq \hat{\pi}_h(\theta)\right) \cap G\right\}. \tag{4.5.11}$$

Moreover,

$$\begin{aligned}\hat{\pi}_h(t) &= (nh^d)^{-1}\int_{\hat{\Theta}} \big\|nh^d(t-u)\big\|_1 L_h\big(u, Y^{(n)}\big)du \\ &= (nh^d)^{-D_b-1}\int_{\hat{U}_n} \big\|nh^d(t-\theta)-u\big\|_1 L_h\big(\theta+u(nh^d)^{-1}, Y^{(n)}\big)du \\ &= (nh^d)^{-D_b-1} L_h\big(\theta, Y^{(n)}\big)\int_{\hat{U}_n} \big\|nh^d(t-\theta)-u\big\|_1 Z_{h,\theta}(u)du.\end{aligned}$$

Hence, $\tau_n = nh^d\big(\hat{\theta}^*(h)-\theta\big)$ is the minimizer of

$$\chi_n(s) = \int_{\hat{U}_n} \big\|s-u\big\|_1 \frac{Z_{h,\theta}(u)}{\int_{\hat{U}_n} Z_{h,\theta}(v)dv}du$$

and we obtain from (4.5.10) and (4.5.11) for any $\varepsilon > 0$

$$\mathbb{P}_f\left(\big\|nh^d(\hat{\theta}^*(h)-\theta)\big\|_1 > \varepsilon, G\right) \le \mathbb{P}_f\left(\inf_{\|s\|_1>\varepsilon}\chi_n(s) \le \chi_n(0), G\right). \tag{4.5.12}$$

Let $0 < r < \varepsilon/3$, be a number whose choice will be done later. We have

$$\chi_n(0) \le r\int_{\hat{U}_n\cap(\|u\|_1\le r)} z_h(u)du + \int_{\hat{U}_n\cap(\|u\|_1>r)} \|u\|_1 z_h(u)du.$$

Note also that

$$\begin{aligned}\inf_{\|s\|_1>\varepsilon}\chi_n(s) &\ge \inf_{\|s\|_1>\varepsilon}\left[\int_{\hat{U}_n\cap(\|u\|_1\le r)}\big(\|s\|_1-\|u\|_1\big)z_h(u)du\right] \\ &\ge (\varepsilon-r)\int_{\hat{U}_n\cap(\|u\|_1\le r)} z_h(u)du.\end{aligned}$$

It yields in particular

$$\begin{aligned}&\chi_n(0) - \inf_{\|s\|_1>\varepsilon}\chi_n(s) \\ &\quad\le -(\varepsilon-2r)\int_{\hat{U}_n\cap(\|u\|_1\le r)} z_h(u)du + \int_{\hat{U}_n\cap(\|u\|_1>r)} \|u\|_1 z_h(u)du.\end{aligned}$$

Thus, $\forall r \in (0, \varepsilon/3)$

$$\begin{aligned}&\mathbb{P}_f\left(\chi_n(0) - \inf_{\|s\|_1>\varepsilon}\chi_n(s) > 0, G\right) \\ &\le \mathbb{P}_f\left(\int_{\hat{U}_n\cap(\|u\|_1>r)} \|u\|_1 z_h(u)du > (\varepsilon-2r)\int_{\hat{U}_n\cap(\|u\|_1\le r)} z_h(u)du, G\right) \\ &\le \mathbb{P}_f\left(\int_{\hat{U}_n\cap(\|u\|_1>r)} \|u\|_1 z_h(u)du > r/2, G\right) \\ &\quad +\mathbb{P}_f\left((\varepsilon-2r)\int_{\hat{U}_n\cap(\|u\|_1\le r)} z_h(u)du < r/2, G\right).\end{aligned} \tag{4.5.13}$$

We note that the second term in (4.5.13) can be control by the first one whenever $0 < r < \varepsilon/3$. Indeed, putting $\hat{U}_n(r) = \hat{U}_n \cap (u \in \mathbb{R}^{D_b} \, : \, \|u\|_1 > r)$ we get

$$\begin{aligned}
&\mathbb{P}_f\left((\varepsilon - 2r)\int_{\hat{U}_n \cap (\|u\|_1 \le r)} z_h(u)du < r/2, G\right) \\
&\le \mathbb{P}_f\left(r\int_{\hat{U}_n} Z_{h,\theta}(v)dv - r\int_{\hat{U}_n(r)} Z_{h,\theta}(u)du < \frac{r}{2}\int_{\hat{U}_n} Z_{h,\theta}(v)dv, G\right) \\
&\le \mathbb{P}_f\left(r\int_{\hat{U}_n(r)} Z_{h,\theta}(u)du > \frac{r}{2}\int_{\hat{U}_n} Z_{h,\theta}(v)dv, G\right) \\
&\le \mathbb{P}_f\left(\int_{\hat{U}_n(r)} \|u\|_1 z_h(u)du > r/2, G\right).
\end{aligned}$$

The last inequality together with (4.5.10), (4.5.12) and (4.5.13) yields

$$\mathbb{P}_f\left(nh^d\big|\hat{f}^h(y) - f(y)\big| \ge \varepsilon, G\right) \le 2\mathbb{P}_f\left(\int_{\hat{U}_n(r)} \|u\|_1 z_h(u)du > \frac{r}{2}, G\right). \tag{4.5.14}$$

Second step: Proof of Assertion 2.
Put $U_n(a) = U_n \cap (u \in \mathbb{R}^{D_b} \, : \, \|u\|_1 > a)$ for all $a > 0$ and $\Gamma_\upsilon = U_n(\upsilon + 1) \setminus U_n(\upsilon)$ for any $\upsilon > a$. Introduce the following notations.

$$\mathcal{I}_\upsilon = \int_{\Gamma_\upsilon} Z_{h,\theta}(u)du, \qquad \mathcal{Q}_\upsilon = \frac{\mathcal{I}_\upsilon}{\int_{\hat{U}_n} Z_{h,\theta}(u)du}.$$

We divide the set Γ_υ into T^{D_b} nonintersecting parts $\Delta_1, \Delta_2, \ldots, \Delta_{T^{D_b}}$ such that

$$\Gamma_\upsilon = \bigcup_{j=1}^{T^{D_b}} \Delta_j, \qquad \Delta_j \cap \Delta_l = \emptyset, \ \forall j, l.$$

The integer number T will be chosen later. Let u_i be the point arbitrary chosen in the set Δ_i. Then

$$S_\upsilon = \sum_i \int_{\Delta_i} Z_{h,\theta}(u_i)du.$$

In fact, we seek to approach the process $\mathcal{I}_\upsilon$ with a process in steps S_υ. Obviously $\bar{\Gamma}_\upsilon = \int_{\Gamma_\upsilon} du$ and we get for any $\sigma > 0$

$$\begin{aligned}
\mathbb{P}_f\big(S_\upsilon > \sigma\big) &\le \mathbb{P}_f\left(\max_i Z_{h,\theta}^{1/2}(u_i)\sqrt{\bar{\Gamma}_\upsilon} > \sqrt{\sigma}\right) \\
&\le \sum_i \mathbb{P}_f\left(Z_{h,\theta}^{1/2}(u_i) > \big(\bar{\Gamma}_\upsilon\big)^{-1/2}\sqrt{\sigma}\right).
\end{aligned}$$

Note that the number of summands on the right-hand side of the last inequality is equal to T^{D_b}. Applying Markov inequality and Lemma 8.2, we obtain

$$\mathbb{P}_f\big(S_\upsilon > \sigma\big) \le T^{D_b}\sqrt{\bar{\Gamma}_\upsilon}\sigma^{-1/2}e^{-g_h(\upsilon)}. \tag{4.5.15}$$

In view of to Lemma 8.1,

$$\mathbb{E}_f\big|S_\upsilon - \mathcal{I}_\upsilon\big| \le \sum_i \int_{\Delta_i} \mathbb{E}_f\big|Z_{h,\theta}(u) - Z_{h,\theta}(u_i)\big|du \le \mathcal{C}_h \sum_i \int_{\Delta_i} \|u - u_i\|_1 du.$$

Note that each summand does not exceed $\big(\upsilon T^{-1}\big)^{D_b+1}$ and, therefore,

$$\mathbb{E}_f\big|S_\upsilon - \mathcal{I}_\upsilon\big| \le \mathcal{C}_h \upsilon^{D_b+1}T^{-1}. \tag{4.5.16}$$

One has

$$\mathbb{P}_f\big(\mathcal{I}_\upsilon > 2\sigma\big) \le \mathbb{P}_f\big(S_\upsilon > \sigma\big) + \mathbb{P}_f\big(\big|S_\upsilon - \mathcal{I}_\upsilon\big| > \sigma\big).$$

Using (4.5.15), (4.5.16) and applying Markov inequality, we get

$$\mathbb{P}_f\big(\mathcal{I}_\upsilon > 2\sigma\big) \le T^{D_b}\sqrt{\bar{\Gamma}_\upsilon}\sigma^{-1/2}e^{-g_h(\upsilon)} + \mathcal{C}_h\upsilon^{D_b+1}T^{-1}\sigma^{-1}. \tag{4.5.17}$$

Note $\mathbb{A} = \Big\{\int_{\hat{U}_n} Z_{h,\theta}(u)du < \frac{\delta^{D_b}}{2}\Big\}$. Since $\mathcal{Q}_\upsilon \le 1$ we obtain for any $\delta > 0$ and $\sigma > 0$

$$\begin{aligned}\mathbb{E}_f\mathcal{Q}_\upsilon &= \mathbb{E}_f\left[\mathcal{Q}_\upsilon\,\mathbb{I}_{\mathbb{A}} + \mathcal{Q}_\upsilon\,\mathbb{I}_{\mathcal{I}_\upsilon > 2\sigma,\mathbb{A}^c} + \mathcal{Q}_\upsilon\mathbb{I}_{\mathcal{I}_\upsilon \le 2\sigma,\mathbb{A}^c}\right]\\ &\le \mathbb{P}_f\left(\int_{\hat{U}_n} Z_{h,\theta}(u)du < \frac{\delta^{D_b}}{2}\right) + \mathbb{P}_f\big(\mathcal{I}_\upsilon > 2\sigma\big) + 4\delta^{-D_b}\sigma.\end{aligned}$$

Using to Lemma 8.3 and the inequality (4.5.17), we have

$$\mathbb{E}_f\mathcal{Q}_\upsilon \le 2\mathcal{C}_h\delta + T^{D_b}\sqrt{\bar{\Gamma}_\upsilon}\sigma^{-1/2}e^{-g_h(\upsilon)} + \mathcal{C}_h\upsilon^{D_b+1}T^{-1}\sigma^{-1} + 4\delta^{-D_b}\sigma.$$

Choosing $T = \Big\lfloor \exp\Big\{\frac{1}{2D_b}g_h(\upsilon)\Big\}\Big\rfloor + 1$, $\sigma = \exp\Big\{-\frac{1}{3D_b}g_h(\upsilon)\Big\}$ and $\delta = \exp\Big\{-\frac{1}{6D_b^2}g_h(a)\Big\}$, we obtain

$$\mathbb{E}_f\mathcal{Q}_\upsilon \le \Big[2\mathcal{C}_h + 2^{D_b}\sqrt{\bar{\Gamma}_\upsilon} + \mathcal{C}_h\upsilon^{D_b+1} + 4\Big]\exp\left\{-\frac{1}{6D_b^2}g_h(a)\right\}.$$

Simplest algebra shows that $\sqrt{\bar{\Gamma}_\upsilon} \le \big(2\upsilon+2\big)^{\frac{D_b}{2}}$, we get

$$\mathbb{E}_f\mathcal{Q}_\upsilon \le \big[\upsilon^{D_b+1} + 2^{D_b}(2\upsilon+2)^{\frac{D_b}{2}} + 6\big]\,\mathcal{C}_h\,\exp\left\{-\frac{1}{6D_b^2}g_h(a)\right\}, \tag{4.5.18}$$

Note that if the event G is realized then $\hat{U}_n(r) \subseteq U_n(r)$ and $U_n(a) = \bigcup_{j=0}^{\infty} \Gamma_{a+j}$. we obtain in view of (4.5.18)

$$\begin{aligned}\mathbb{E}_f \int_{\hat{U}_n \cap (\|u\|_1 > a)} \|u\|_1 z_h(u)\, \mathbb{I}_G\, du &\leq \sum_{j=0}^{\infty} \big(a+j+1\big)\mathbb{E}_f \mathcal{Q}_{a+j} \\ &= \Sigma^*\big(M(f)\big)\, a\, \mathcal{B}_a\, \mathcal{C}_h\, \exp\left\{-\frac{1}{6D_b^2} g_h(a)\right\}.\end{aligned}$$

where we have put $\mathcal{B}_a = a^{D_b+1} + 2^{D_b}(2a+2)^{\frac{D_b}{2}} + 6 + D_b\left(a^{D_b} + (2a+2)^{\frac{D_b}{2}-1}\right)$.

Third step: Conclusion.
In view of Assertion 2, choosing $r = \varepsilon/4$ we get

$$\mathbb{E}_f \int_{\hat{U}_n \cap (\|u\|_1 > \varepsilon/4)} \|u\|_1 z_h(u)\, \mathbb{I}_G\, du \leq \frac{\varepsilon}{4}\, \Sigma^*\big(M(f)\big)\, \mathcal{B}_{\varepsilon/4}\, \mathcal{C}_h\, e^{-\frac{1}{6D_b^2} g_h(\varepsilon/4)}. \tag{4.5.19}$$

Using the Markov inequality, we have in view of (4.5.19)

$$\mathbb{P}_f\left(\int_{\hat{U}_n \cap (\|u\|_1 > \varepsilon/4)} \|u\|_1 z_h(u) du > \frac{\varepsilon}{8},\, G\right) \leq 2\Sigma^*\big(M(f)\big)\, \mathcal{B}_{\varepsilon/4}\, \mathcal{C}_h\, e^{-\frac{1}{6D_b^2} g_h(\varepsilon/4)}.$$

The assertion of Pro follows now from the last inequality, Assertion 1 and the definitions of $\mathcal{C}_h, g_h(\cdot)$ and the function $\mathfrak{B}(\cdot,\cdot)$. ∎

4.5.3 Proof of Proposition 5

To prove the proposition it suffice to integrate the inequality obtained in Proposition 4 and to use the following lemma which will be extensively exploited in the sequel.

Lemma 9. *There exists $\lambda > 0$ such that $\forall n > 1$ and $\forall h \in \mathcal{H}_n$, we have*

$$\lambda_n(h) \geq \lambda.$$

where $\lambda_n(h)$ is the smallest eigenvalue of the matrix

$$\mathcal{M}_{nh}(y) = \frac{1}{nh^d} \sum_{i=1}^{n} K^\top\left(\frac{X_i - y}{h}\right) K\left(\frac{X_i - y}{h}\right) \mathbb{I}_{V_h(y)}(X_i),$$

and $K(z)$ is the D_b-dimensional vector of the monomials z^p, $p \in \mathcal{P}_b$.

Proof of Proposition 5. In order to simplify the proof, let us introduce the following constants

$$c_1 = \frac{(1+6D_b^2)}{6A(f)D_b^2}, \quad c_2 = \frac{\lambda}{432M(f)\,D_b^3}. \tag{4.5.20}$$

By definition of $A(f)$, $M(f)$, $\mathcal{B}(.,.)$ respectively in (4.1.11), (4.5.5) and A, M, we have the following inequality $\mathfrak{B}\big(A(f),M(f)\big) \leq \mathfrak{B}(A,M)$. By integration of Proposition 4 and using Lemma 9, we get for any $q \geq 1$ and $f \in \mathbb{H}_d(\beta,L,M,A)$

$$\begin{aligned}
&\mathbb{E}_f\big|\hat{f}^h(y) - f(y)\big|^q \mathbb{I}_G \\
&= \int_0^{+\infty} \eta^{q-1}\mathbb{P}_f\left(\big|\hat{f}^h(y) - f(y)\big| \geq \eta, G\right) d\eta \\
&= (nh^d)^{-q} \int_0^{+\infty} \eta^{q-1}\mathbb{P}_f\left(\big|\hat{f}^h(y) - f(y)\big| \geq \frac{\eta}{nh^d}, G\right) d\eta \\
&= (nh^d)^{-q} \left[\int_0^{\frac{c_1}{c_2}\mathcal{N}_h} \eta^{q-1}\, d\eta \right. \\
&\qquad \left. + \int_{\frac{c_1}{c_2}\mathcal{N}_h}^{+\infty} \eta^{q-1}\mathbb{P}_f\left(\big|\hat{f}^h(y) - f(y)\big| \geq \frac{\eta}{nh^d}, G\right) d\eta\right] \\
&\leq (nh^d)^{-q} \left[\frac{c_1^q}{q\,c_2^q}\mathcal{N}_h^q + \int_{\frac{c_1}{c_2}\mathcal{N}_h}^{+\infty} \eta^{q-1}\mathfrak{B}(A,M)\,\mathcal{E}(h)\,e^{-c_2\eta}\,d\eta\right]. \qquad (4.5.21)
\end{aligned}$$

In seeking this type of result (Control of the risk), the new point is to control the term $\mathcal{E}(h)$ which can be very large compared to the rate $(nh^d)^{-1}$. It is a consequence that the bayesian estimator is non-linear. The bias-variance ratio $\mathcal{N}_h$ (defined in (4.5.4)), which appears in $\mathcal{E}(h)$, allows to obtain similar results with linear estimators.

By calculation of the integral in (4.5.21), b_h and $\mathcal{N}_h$ respectively defined in (4.5.3) and (4.5.4), the assertion of Proposition 5 is proved.

∎

4.5.4 Proof of Theorem 11

By definition of $\bar{h} = (Ln)^{-\frac{1}{\beta+d}}$, we have

$$Ldn\bar{h}^{\beta+d} = d, \quad (n\bar{h}^d)^{-q} = L^{\frac{qd}{\beta+d}}\varphi_n^q(\beta). \tag{4.5.22}$$

We take $\hat{M} = M$ and $\hat{A} = A$ (known in minimax case), by definition of $\bar{f}^{\bar{h}}$ in (4.1.7), (4.1.8), therefore Proposition 5 can be obtained by replacing $\hat{f}$ by $\bar{f}$. Indeed the event G, defined in Section 4.5.1, is not necessary when the set of coefficients $\Theta(A,M)$, in the definition of

$\bar{f}$ (see (4.1.6)), is not random. In this case, the reasoning of proofs of Propositions 4 and 5 remains true for the estimator $\bar{f}$, then Proposition 5 can be applied for the estimator $\bar{f}$ with same constants. Using (4.5.22), we obtain

$$\mathbb{E}_f\big|\bar{f}^{\bar{h}}(y) - f(y)\big|^q \le C_q^* \, d \, L^{\frac{qd}{\beta+d}} \varphi_n^q(\beta).$$

where λ is defined in the Lemma 5 in Section 4.5.3. The theorem 11 is proved.

∎

4.5.5 Proof of Theorem 14

This Proof is based on the Lepski scheme developed by Lepski [1991] and adapted for the bandwidth selection by Lepski, Mammen, and Spokoiny [1997]. We start the proof with formulating auxiliary Lemmas whose proofs are given in Appendix (Section 4.7). Define $J_1 = Ld(1 + 6D_b^2)/6A(f)D_b^2$ and

$$h^* = \left[\frac{c(1 + (b-\beta)\ln n)}{n}\right]^{\frac{1}{\beta+d}}, \quad c < \frac{1 \wedge qdJ_1^{-1}}{(b+d)(\beta+d)}.$$

and let the integer κ be defined as follows.

$$2^{-\kappa} h_{\max} \le h^* < 2^{-\kappa+1} h_{\max}. \tag{4.5.23}$$

The definitions of h^* and κ (4.5.23) imply the following Lemmas.

Lemma 10.

$$\mathbb{E}_f\big|\hat{f}^{(k)}(y) - f(y)\big|^q \mathbb{I}_G \le \bar{C}_q \frac{(1 + k\ln 2)^q}{(nh_k^d)^q}, \quad \forall k \ge \kappa,$$

where $\bar{C}_q = C_q^* \dfrac{(\beta+d)(1 \wedge qdJ_1^{-1})}{(\beta+d-1)(Ld)^{-1}}$.

Lemma 11. *For any* $f \in \mathbb{H}_d(\beta, L, M, A)$ *and any* $k \ge \kappa + 1$

$$\mathbb{P}_f\big(\hat{k} = k,\ G\big) \le J_2 \mathfrak{B}(A, M) \exp\left\{J_1 n (h^*)^{\beta+d}\right\} 2^{-(k-1)(8qd+4)},$$

where $J_2 = (1 - 2^{-(8qd+4)})^{-1}$.

Lemma 12. *There exists a universal constant* $\vartheta > 0$ *such that*

$$\limsup_{n\to\infty} \sup_{f \in \mathbb{H}_d(\beta,L,M,A)} \exp\left\{\frac{An^{\frac{b}{b+d}}}{16M\vartheta^2 D_b^2}\right\} \mathbb{P}_f\big(G^c\big) = 0.$$

Proof of Theorem 14. We decompose the risk as follows

$$\begin{aligned}&\mathbb{E}_f\big|\hat f^{(\hat k)}(y)-f(y)\big|^q\mathbb{I}_G\\&\quad\le \mathbb{E}_f\big|\hat f^{(\hat k)}(y)-f(y)\big|^q\mathbb{I}_{\hat k\le\kappa,G}+\mathbb{E}_f\big|\hat f^{(\hat k)}(y)-f(y)\big|^q\mathbb{I}_{\hat k>\kappa,G}\\&\quad= R_1(f)+R_2(f).\end{aligned}\tag{4.5.24}$$

First we control R_1. Obviously

$$\big|\hat f^{(\hat k)}(y)-f(y)\big|\le\big|\hat f^{(\hat k)}(y)-\hat f^{(\kappa)}(y)\big|+\big|\hat f^{(\kappa)}(y)-f(y)\big|.$$

Note that the realization of the event G implies $\hat M\le 3M(f)/2$. This together with the definition of $\hat k$ yields

$$\big|\hat f^{(\hat k)}(y)-\hat f^{(\kappa)}(y)\big|\mathbb{I}_{\hat k\le\kappa,G}\le Cs_n(\kappa),\quad s_n(k)=(1+k\ln 2)^q(nh_k^d)^{-q},$$

where $C=288MD_b^3\lambda^{-1}(32qd+16)$. In view of Lemma 10 we also get

$$\mathbb{E}_f\big|\hat f^{(\kappa)}(y)-f(y)\big|^q\le\bar C_q s_n(\kappa).$$

Noting that the right hand side of the obtain inequality is independent of f and taking into account the definition of κ and h^* we obtain

$$\limsup_{n\to\infty}\sup_{f\in\mathbb{H}_d(\beta,L,A,M)}\phi_n^{-q}(\beta)R_1(f)<\infty.\tag{4.5.25}$$

Now let us bounded from above R_2. Applying Cauchy-Schwartz inequality we have in view of Lemma 11

$$\begin{aligned}R_2(f)&=\sum_{k>\kappa}^{\mathrm{k}_n}\mathbb{E}_f\big|\hat f^{(k)}(y)-f(y)\big|^q\mathbb{I}_{\left[\hat k=k,G\right]}\\&\le\sum_{k>\kappa}\big(\mathbb{E}_f\big|\hat f^{(k)}(y)-f(y)\big|^{2q}\big)^{1/2}\sqrt{\mathbb{P}_f\big\{\hat k=k,G\big\}}\\&=\Delta(h^*)\sum_{k>\kappa}\big(\mathbb{E}_f\big|\hat f^{(k)}(y)-f(y)\big|^{2q}\big)^{1/2}2^{-(k-1)(4qd+2)},\end{aligned}\tag{4.5.26}$$

where we have put $\Delta(h^*)=J_2\mathfrak{B}(A,M)\exp\big\{J_1n(h^*)^{\beta+d}\big\}$. We obtain from Lemma 10 and (4.5.26)

$$R_2(f)\le\ J_3\,(nh_{\max}^d)^{-q}\exp\big\{J_1n(h^*)^{\beta+d}\big\},\tag{4.5.27}$$

where

$$J_3=J_2\mathfrak{B}(A,M)\,2^{4qd+2}\,\bar C_{2q}^{1/2}\sum_{s\ge0}(1+s\ln 2)^q2^{-3sdq-2}.$$

It remains to note that the definition of h^* implies that

$$\limsup_{n\to\infty} \phi_n^{-q}(\beta)(nh_{\max}^d)^{-q} \exp\left\{J_1 n (h^*)^{\beta+d}\right\} < \infty$$

and that the right hand side of (4.5.27) is independent of f. Thus,we have

$$\limsup_{n\to\infty} \sup_{f\in\mathbb{H}_d(\beta,L,A,M)} \phi_n^{-q}(\beta) R_2(f) < \infty.$$

that yields together with (4.5.24) and (4.5.25)

$$\limsup_{n\to\infty} \sup_{f\in\mathbb{H}_d(\beta,L,A,M)} \phi_n^{-q}(\beta)\mathbb{E}_f\big|\hat{f}^{(\hat{k})}(y) - f(y)\big|^q \mathbb{I}_G < \infty.$$

To get the assertion of the theorem it suffices to show that

$$\limsup_{n\to\infty} \sup_{f\in\mathbb{H}_d(\beta,L,A,M)} \phi_n^{-q}(\beta)\mathbb{E}_f\big|\hat{f}^{(\hat{k})}(y) - f(y)\big|^q \mathbb{I}_{G^c} < \infty. \tag{4.5.28}$$

Note that $\hat{f}^{(\hat{k})}(y) \le 4\hat{M}$ in view of (4.3.6). Note also that the local least square estimator $\tilde{\delta}$ is linear function of observation $Y^{(n)}$ and, moreover $0 \le Y_i \le M, i = 1, ..., n$. This together with the definition of $\hat{M}$, (expression (4.3.5)) allows us to state that there exist $0 < J_4 < +\infty$ such that $\big|\hat{f}^{(\hat{k})}(y) - f(y)\big| \le J_4 M$. Here we also have taken into account that $||f||_\infty \le M$.

Finally we obtain

$$\mathbb{E}_f\big|\hat{f}^{(\hat{k})}(y) - f(y)\big|^q \mathbb{I}_{G^c} \le J_4^q\, M^q \mathbb{P}_f\{G^c\}.$$

and (4.5.28) follows now from Lemma 11. ∎

4.6 Proofs of lower bounds

The proofs of Theorems 10 and 13 are based on the following proposition.

Put $\phi_n(\gamma) = \big[n^{-1}\big(1 + (b-\gamma)\ln n\big)\big]^{\frac{\gamma}{\gamma+d}}$, $\gamma \in (0, b]$ and let

$$R_n^{(q)}(\tilde{f}, v) = \sup_{f\in\mathbb{H}_d(\alpha,L,M,A)} \mathbb{E}_f\left[\phi_n^{-q}(\alpha)|\tilde{f}(y) - f(y)|^q\right] + \sup_{f\in\mathbb{H}_d(\beta,L,M,A)} \mathbb{E}_f\left[n^{-vq}\phi_n^{-q}(\beta)|\tilde{f}(y) - f(y)|^q\right].$$

where $v \ge 0$ and $\alpha, \beta \in (0, b]^2$.

Proposition 6. *Let Ψ be admissible family of normalizations such that*

$$\psi_n(\alpha)/\phi_n(\alpha) \xrightarrow[n\to\infty]{} 0.$$

Then, for any $0 \le v < (\beta-\alpha)/(\beta+1)(\alpha+1)$

$$\liminf_{n\to\infty} \inf_{\tilde{f}} R_n^{(q)}(\tilde{f}, v) > 0.$$

The proof is given in section 4.6.3.

4.6.1 Proof of Theorem 10

Using the proposition 6 for $\beta = \alpha$, we have to choose $v = 0$ and one gets

$$\begin{aligned} R_{n,q}\big[\mathbb{H}_d(\beta, L, M, A)\big] &= R_n^{(q)}(\tilde{f}, 0) \\ &= \sup_{f \in \mathbb{H}_d(\alpha, L, M, A)} \mathbb{E}_f \left[n^{-q\frac{\alpha}{\alpha+1}} \big|\tilde{f}(y) - f(y)\big|^q \right] > 0, \quad \forall \tilde{f}. \end{aligned}$$

∎

4.6.2 Proof of Theorem 13

I. To proof of the first assertion of the theorem it suffices to consider the family $\{\upsilon_n(\beta)\}_{\beta \in (0,b]}$, where $\upsilon_n(\alpha) = \varphi_n(\alpha)$ and $\upsilon_n(\beta) = 1$ for any $\beta \neq \alpha$. The corresponding attainable estimator is the estimator being minimax on $\mathbb{H}_d(\alpha, L, M, A)$.

II. Let us consider the family $\{\phi_n(\beta)\}_{\beta \in]0,b]}$, which is admissible in view of Theorem 14. First, we note that $\gamma = b$ is not possible since $\phi_n(b) = \varphi_n(b)$ the minimax rate of convergence on $\mathbb{H}_d(b, L, M, A)$.

Thus we assume that γ satisfying (4.3.2) belongs to $\delta \in]0, b[$. Let $\hat{f}^\Psi$ be a $\Psi^{(n)}$-attainable estimator. Since $\psi_n(\alpha)/\phi_n(\alpha) \to 0, n \to \infty$ in view of (4.3.2) then obviously

$$\limsup_{n\to\infty} \sup_{f \in \mathbb{H}_d(\gamma, L, M, A)} \mathbb{E}_f \left[\phi_n^{-q}(\gamma) \big|\hat{f}^\Psi(y) - f(y)\big|^q \right] = 0.$$

Therefore, applying Proposition 6 with $v = 0$ we have for any $\beta < \gamma$

$$\limsup_{n\to\infty} \sup_{f \in \mathbb{H}_d(\beta, L, M, A)} \mathbb{E}_f \left[\phi_n^{-q}(\beta) \big|\hat{f}^\Psi(y) - f(y)\big|^q \right] > 0.$$

We conclude that necessarily $\psi_n(\beta) \gtrsim \phi_n(\beta)$ for any $\beta < \gamma$.

Moreover for any$\beta > \gamma$ applying Proposition 6 with an arbitrary $0 \leq v < (\beta - \gamma)/(\beta + 1)(\gamma + 1)$ we obtain that

$$\psi_n(\beta) \gtrsim n^v \phi_n(\beta), \ \ \beta > \gamma.$$

It remains to note that the form of rate of convergence proved in Theorem 10 implies that

$$\phi_n(\gamma)/\psi_n(\gamma) = o\left([\ln n]^{\frac{\gamma}{\gamma+d}} \right).$$

∎

4.6.3 Proof of Proposition 6

Let $\varkappa > 0$ the parameter whose choice will be done later. Put

$$h = \left(\varkappa \frac{1 + (\beta - \alpha)\ln n}{n}\right)^{\frac{1}{\alpha+d}}.$$

Later on without loss of generality we will assume that $L > 1$.

Consider the functions: $f_0 \equiv 1$ and

$$f_1(x) = 1 - \big(L-1\big)\varkappa^{\frac{\alpha}{\alpha+d}}\phi_n(\alpha)F\left(\frac{x_1 - y_1}{h}, ..., \frac{x_d - y_d}{h}\right), \quad x \in [0,1]^d.$$

Here F is a compactly supported positive function belonging to $\mathbb{H}_d(\alpha, 1, M, A)$ such that $F(0) = 1 = \max_x F(x)$.

It is easily seen that $f_1 \in \mathbb{H}_d(\alpha, L, M, A)$. Therefore, we have

$$\begin{aligned} R_n^{(q)}(\tilde f, v) &\geq \mathbb{E}_0\left|n^{-v}\phi_n^{-1}(\beta)\big(\tilde f(y) - 1\big)\right|^q + \mathbb{E}_1\left|\phi_n^{-1}(\alpha)\big(\tilde f(y) - f_1(y)\big)\right|^q \\ &\geq \mathbb{E}_0\left|n^{-v}\phi_n^{-1}(\beta)\big(\tilde f(y) - 1\big)\right|^q + \mathbb{E}_1\left|\phi_n^{-1}(\alpha)\big(\tilde f(y) - 1\big) + z\right|^q, \end{aligned}$$

where $z = \big(L-1\big)\varkappa^{\frac{\alpha}{\alpha+1}}F(0)$. Set

$$\tilde\lambda = \phi_n^{-1}(\alpha)\big(1 - \tilde f(y)\big), \quad \varsigma_n = n^{-v}\frac{\phi_n(\alpha)}{\phi_n(\beta)} = n^{-v}\left(\frac{\ln n}{n}\right)^{-\varrho},$$

where $\varrho = \frac{\beta-\alpha}{(\beta+1)(\alpha+1)}$. We get

$$\begin{aligned} R_n^{(q)}(\tilde f, v) &\geq \mathbb{E}_0\big|\varsigma_n\tilde\lambda\big|^q + \mathbb{E}_1\big|z - \tilde\lambda\big|^q \\ &\geq \mathbb{E}_0\big|\varsigma_n\tilde\lambda\big|^q\mathbb{I}_{\{|\tilde\lambda| > z/2\}} + \mathbb{E}_1\big|z - \tilde\lambda\big|^q\mathbb{I}_{\{|\tilde\lambda| \leq z/2\}} \\ &\geq \mathbb{E}_0\big|\varsigma_n\frac{z}{2}\big|^q\mathbb{I}_{\{|\tilde\lambda| > z/2\}} + \mathbb{E}_1\big|\frac{z}{2}\big|^q\mathbb{I}_{\{|\tilde\lambda| \leq z/2\}}. \end{aligned} \tag{4.6.1}$$

Noting that $f_1 \leq f_0$, since F is positive, and putting $c_n\big(Y^{(n)}\big) = \mathbb{I}_{\{|\tilde\lambda| > z/2\}}$ we obtain

$$\begin{aligned} R_n^{(q)}(\tilde f, v) \geq\ & \varsigma_n^q\ \frac{z^q}{2^q}\frac{\prod_{i=1}^n f_1(X_i)}{\prod_{i=1}^n f_1(X_i)}\int_0^{f_1(X_1)}\dots\int_0^{f_1(X_n)} c_n(x)dx_1\dots dx_n \\ &+ \frac{z^q}{2^q}\frac{1}{\prod_{i=1}^n f_1(X_i)}\int_0^{f_1(X_1)}\dots\int_0^{f_1(X_n)} 1 - c_n(x)dx_1\dots dx_n. \end{aligned} \tag{4.6.2}$$

We have

$$\begin{aligned} \prod_{i=1}^n f_1(X_i) &= \prod_{i=1}^n\left(1 - (L-1)\varkappa^{\frac{\alpha}{\alpha+d}}\phi_n(\alpha)F\left(\frac{X_i - y}{h}\right)\right) \\ &\geq \Big(1 - (L-1)\varkappa^{\frac{\alpha}{\alpha+d}}\phi_n(\alpha)\Big)^{nh^d} \geq e^{-(L-1)\varkappa}n^{-(L-1)\varkappa(\beta-\alpha)}. \end{aligned} \tag{4.6.3}$$

We obtain in view of (4.6.2) and (4.6.3)

$$\begin{aligned} R_n^{(q)}(\tilde{f},v) &\geq \varsigma_n^q \frac{z^q}{2^q} e^{-(L-1)\varkappa} n^{-(L-1)\varkappa(\beta-\alpha)} \\ &\quad \times \frac{1}{\prod_{i=1}^n f_1(X_i)} \int_0^{f_1(X_1)} \dots \int_0^{f_1(X_n)} c_n(x) dx_1 \dots dx_n \\ &\quad + \frac{z^q}{2^q} \frac{1}{\prod_{i=1}^n f_1(X_i)} \int_0^{f_1(X_1)} \dots \int_0^{f_1(X_n)} 1 - c_n(x) dx_1 \dots dx_n \\ &\geq \frac{z^q}{2^q} \left(1 \wedge \varsigma_n^q e^{-(L-1)\varkappa} n^{-(L-1)\varkappa(\beta-\alpha)}\right). \end{aligned}$$

<u>*Case 1:*</u> $\beta = \alpha$. Choosing $\varkappa = 1$, and noting that $\varsigma_n = 1$ and $\prod_{i=1}^n f_1(X_i) \geq e^{-(L-1)}$, we deduce from (4.6.2) that yields:

$$\inf_{\tilde{f}} R_n^{(q)}(\tilde{f},v) \geq \frac{(L-1)^q}{2^q} e^{-(L-1)} > 0.$$

<u>*Case 2:*</u> $\beta > \alpha$. Put

$$\varkappa = \frac{q\big(\varrho - v\big) - t_n}{1 + (L-1)(\beta-\alpha)} > 0, \;\; t_n = \frac{q}{\ln n} \ln \frac{1}{\big(1 + (\beta-\alpha)\ln n\big)^{-\varrho}} \xrightarrow[n\to\infty]{} 0.$$

This choice provides us with the following bound

$$\begin{aligned} \varsigma_n^q e^{-(L-1)\varkappa} n^{-(L-1)\varkappa(\beta-\alpha)} &= \big(1 + (\beta-\alpha)\ln n\big)^{-q\varrho} e^{-(L-1)\varkappa} n^{q(\varrho-v)-(L-1)\varkappa(\beta-\alpha)} \\ &\geq \big(1 + (\beta-\alpha)\ln n\big)^{-q\varrho} e^{-\frac{2}{3}q(L-1)} n^{t_n} \geq e^{-\frac{2}{3}q(L-1)}. \end{aligned}$$

This yields

$$\inf_{\tilde{f}} R_n^{(q)}(\tilde{f},v) \geq \frac{(L-1)^q \varkappa^{q\frac{\alpha}{\alpha+1}}}{2^q} e^{-\frac{2}{3}q(L-1)} > 0.$$

∎

4.7 Appendix

Proof of Lemma 8 Later on without loss generality we will suppose that $nh^d \in \mathbb{N}^*$. In order to simplify understanding of this proof, we note the approximation polynomial $\mathcal{A}_u^i = f_{\theta+u(nh^d)^{-1}}(X_i)$, $i = 1, \ldots, n$ for all $u \in U_n$.

1. Note that for $u \in U_n$

$$\mathbb{E}_f Z_{h,\theta}(u) \leq \prod_{i:\, X_i \in V_h(y)} \frac{\mathcal{A}_0^i}{f(X_i)} \leq e^{\mathcal{N}_h/A(f)}. \tag{4.7.1}$$

This result is the consequence of the definition of $Z_{h,\theta}$ in (4.5.9) and the following calculation

$$\mathbb{E}_f \mathbb{I}_{[Y_i \leq \mathcal{A}_u^i]} = \mathbb{P}_f\left(Y_i \leq \mathcal{A}_u^i\right) = 1 \wedge \frac{\mathcal{A}_u^i}{f(X_i)}.$$

In (4.7.1), the second inequality is obtained with classical inequality $1 + \rho \leq e^\rho, \rho \in \mathbb{R}$ and recall that $f_\theta(x) \geq f(x)$.

$$\prod_{i:\, X_i \in V_h(y)} \frac{\mathcal{A}_0^i}{f(X_i)} = \prod_{i:\, X_i \in V_h(y)} \left(1 + \frac{\mathcal{A}_0^i - f(X_i)}{f(X_i)}\right) \leq \exp\left\{b_h \times nh^d/A(f)\right\}$$

Case 1: If $\|u_1 - u_2\|_1 \geq 1$, the inequality (4.7.1) allows to get

$$\mathbb{E}_f\left|Z_{h,\theta}(u_1) - Z_{h,\theta}(u_2)\right| \leq \mathbb{E}_f Z_{h,\theta}(u_1) + \mathbb{E}_f Z_{h,\theta}(u_2) \leq 2e^{\mathcal{N}_h/A(f)}\|u_1 - u_2\|_1.$$

Case 2: Assume now that $\|u_1 - u_2\|_1 < 1$ and introduce the random events

$$\begin{aligned} F_1 &= \left\{\forall i = 1, \ldots, n:\ Y_i \leq \mathcal{A}_{u_1}^i \wedge \mathcal{A}_{u_2}^i\right\}, \\ F_2 &= \left\{\forall i = 1, \ldots, n:\ Y_i \leq \mathcal{A}_{u_1}^i \vee \mathcal{A}_{u_2}^i\right\} \\ &\quad \cap \left\{\exists i:\ Y_i > \mathcal{A}_{u_1}^i \wedge \mathcal{A}_{u_2}^i\right\}, \\ F_3 &= \left\{\exists i:\ Y_i > \mathcal{A}_{u_1}^i \vee \mathcal{A}_{u_2}^i\right\}. \end{aligned}$$

We have used the following notations: $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$, $a, b \in \mathbb{R}$. For any $(u_1, u_2) \in U_n^2$, we have

$$\begin{aligned} &\mathbb{E}_f\left|Z_{h,\theta}(u_1) - Z_{h,\theta}(u_2)\right| = \mathbb{E}_f\left|Z_{h,\theta}(u_1) - Z_{h,\theta}(u_2)\right| \mathbb{I}_{[F_1]} \\ &+\mathbb{E}_f\left|Z_{h,\theta}(u_1) - Z_{h,\theta}(u_2)\right| \mathbb{I}_{[F_2]} + \mathbb{E}_f\left|Z_{h,\theta}(u_1) - Z_{h,\theta}(u_2)\right| \mathbb{I}_{[F_3]} \\ &= \mathcal{K}_1 + \mathcal{K}_2 + \mathcal{K}_3. \end{aligned}$$

The following bound will be extensively exploited in the sequel.

$$f_v(x) \geq 2v_{0,\ldots,0} - ||v||_1 \geq 0.25A(f), \quad \forall v \in \Theta(A(f)/4, 9M(f)),\ x \in [0,1]^d.$$

Control of $\mathcal{K}_1$.

$$\mathcal{K}_1 = \left| \prod_{i:\, X_i \in V_h(y)} \frac{\mathcal{A}_0^i}{\mathcal{A}_{u_1}^i} - \prod_{i:\, X_i \in V_h(y)} \frac{\mathcal{A}_0^i}{\mathcal{A}_{u_2}^i} \right| \mathbb{P}_f\{F_1\}, \tag{4.7.2}$$

and

$$\mathbb{P}_f\{F_1\} = \prod_{i:\, X_i \in V_h(y)} \mathbb{P}_f\{Y_i \le \mathcal{A}_{u_1}^i \wedge \mathcal{A}_{u_2}^i\} \le \prod_{i:\, X_i \in V_h(y)} \frac{\mathcal{A}_{u_1}^i \wedge \mathcal{A}_{u_2}^i}{f(X_i)}. \tag{4.7.3}$$

Therefore,

$$\begin{aligned} \mathcal{K}_1 &\le \left(1 - \prod_{i:\, X_i \in V_h(y)} \frac{\mathcal{A}_{u_1}^i \wedge \mathcal{A}_{u_2}^i}{\mathcal{A}_{u_1}^i \vee \mathcal{A}_{u_2}^i}\right) \prod_{i:\, X_i \in V_h(y)} \frac{\mathcal{A}_0^i}{f(X_i)} \\ &\le e^{\mathcal{N}_h/A(f)} \left(1 - \exp\left\{ \sum_{i:\, X_i \in V_h(y)} \ln \frac{\mathcal{A}_{u_1}^i \wedge \mathcal{A}_{u_2}^i}{\mathcal{A}_{u_1}^i \vee \mathcal{A}_{u_2}^i} \right\}\right). \end{aligned} \tag{4.7.4}$$

Remember that $\left|\mathcal{A}_{u_1}^i - \mathcal{A}_{u_2}^i\right| \le (nh^d)^{-1}\|u_1 - u_2\|_1$ and $\mathcal{A}_u^i \ge A(f)/4$. Let us give the following calculation with inequality of finite increments

$$\ln \frac{\mathcal{A}_{u_1}^i \wedge \mathcal{A}_{u_2}^i}{\mathcal{A}_{u_1}^i \vee \mathcal{A}_{u_2}^i} = -\left|\ln \mathcal{A}_{u_1}^i \wedge \mathcal{A}_{u_2}^i - \ln \mathcal{A}_{u_1}^i \vee \mathcal{A}_{u_2}^i\right| \ge -\frac{(nh^d)^{-1}\|u_1 - u_2\|_1}{\mathcal{A}_{u_1}^i \wedge \mathcal{A}_{u_2}^i}$$

Using last inequalities, (4.7.2), (4.7.3), (4.7.4) and the well known inequality $1 - e^{-\rho} \le \rho$, we have

$$\mathcal{K}_1 \le \frac{1}{A(f)} e^{\mathcal{N}_h/A(f)} \|u_1 - u_2\|_1.$$

Control of $\mathcal{K}_2$. Put

$$\begin{aligned} F_2 &= \left\{\forall i = 1, \dots, n:\ Y_i \le \mathcal{A}_{u_1}^i \vee \mathcal{A}_{u_2}^i\right\} \\ &\qquad \setminus \left\{\forall i = 1, \dots, n:\ Y_i \le \mathcal{A}_{u_1}^i \wedge \mathcal{A}_{u_2}^i\right\} \\ &= G \setminus F_1. \end{aligned}$$

and define

$$\begin{aligned} \mathcal{G}_1 &= \left\{X_i \in V_h(y):\ \mathcal{A}_{u_1}^i \vee \mathcal{A}_{u_2}^i < f(X_i)\right\}, \\ \mathcal{G}_2 &= \left\{X_i \in V_h(y):\ \mathcal{A}_{u_1}^i \wedge \mathcal{A}_{u_2}^i < f(X_i)\right\}. \end{aligned}$$

Note that $F_1 \subseteq G$ and, therefore,

$$\begin{aligned}\mathcal{K}_2 &\leq \prod_{i:\, X_i \in V_h(y)} \frac{\mathcal{A}_0^i}{\mathcal{A}_{u_1}^i \wedge \mathcal{A}_{u_2}^i} \left(\mathbb{P}_f\{G\} - \mathbb{P}_f\{F_1\}\right) \\ &= \prod_{i:\, X_i \in V_h(y)} \frac{\mathcal{A}_0^i}{\mathcal{A}_{u_1}^i \wedge \mathcal{A}_{u_2}^i} \left(\prod_{i:\, X_i \in \mathcal{G}_1} \frac{\mathcal{A}_{u_1}^i \vee \mathcal{A}_{u_2}^i}{f(X_i)} - \prod_{i:\, X_i \in \mathcal{G}_2} \frac{\mathcal{A}_{u_1}^i \wedge \mathcal{A}_{u_2}^i}{f(X_i)} \right)\end{aligned}$$

The definition of $\mathcal{G}_2$ implies

$$\prod_{i:\, X_i \in V_h(y)} \frac{1}{\mathcal{A}_{u_1}^i \wedge \mathcal{A}_{u_2}^i} \leq \prod_{i:\, X_i \in \mathcal{G}_2} \frac{1}{\mathcal{A}_{u_1}^i \wedge \mathcal{A}_{u_2}^i} \prod_{i:\, X_i \in \mathcal{G}_2^c} \frac{1}{f(X_i)}$$

Since $\mathcal{G}_1 \subseteq \mathcal{G}_2$, $\|u_1 - u_2\|_1 < 1$ and $|f_u(x)| \leq ||u||_1$, $\forall x \in [0,1]^d$, $\forall u \in U_n$, using the last inequality, we obtain

$$\begin{aligned}\mathcal{K}_2 &\leq \prod_{i:\, X_i \in V_h(y)} \frac{\mathcal{A}_0^i}{f(X_i)} \left(\prod_{i:\, X_i \in \mathcal{G}_2} \frac{\mathcal{A}_{u_1}^i \vee \mathcal{A}_{u_2}^i}{\mathcal{A}_{u_1}^i \wedge \mathcal{A}_{u_2}^i} - 1 \right) \\ &\leq 4 D_b e^{1+\mathcal{N}_h/A(f)} \|u_1 - u_2\|_1 / A(f).\end{aligned}$$

It remains to note that $\mathcal{K}_3 = 0$ and the first assertion of the lemma is proved with the bounds of $\mathcal{K}_1$ and $\mathcal{K}_2$.

2. For any $u \in U_n$, since the random variables $(Y_i)_i$ are independent we have,

$$\mathbb{E}_f Z_{n,\theta}^{1/2}(u) = \prod_{i:\, X_i \in V_h(y)} \sqrt{\frac{\mathcal{A}_0^i}{\mathcal{A}_u^i}}\, \mathbb{P}_f\left\{Y_i \leq \mathcal{A}_u^i\right\}.$$

For any i, we have

$$\sqrt{\frac{\mathcal{A}_0^i}{\mathcal{A}_u^i}}\, \mathbb{P}_f\left\{Y_i \leq \mathcal{A}_u^i\right\} = \sqrt{\frac{\mathcal{A}_0^i}{\mathcal{A}_u^i}} \left[1 \wedge \frac{\mathcal{A}_u^i}{f(X_i)}\right] \leq \frac{\mathcal{A}_0^i}{f(X_i)} \left[\frac{f(X_i)}{\sqrt{\mathcal{A}_0^i}\sqrt{\mathcal{A}_u^i}} \wedge \frac{\sqrt{\mathcal{A}_u^i}}{\sqrt{\mathcal{A}_0^i}}\right].$$

Remind that in view of (4.5.8) $f_\theta(x) \geq f(x)$ and $0 < f_\theta(x) \leq 3M(f)$ for $x \in V_h(y)$.

Moreover, for $u \in U_n = nh^d\big(\Theta\big(A(f)/4, 9M(f)\big) - \theta\big)$, $0 < f_{\theta + u(nh^d)^{-1}}(x) \leq 9M(f)$. Thus for all $i : X_i \in V_h(y)$,

$$\sqrt{\frac{\mathcal{A}_0^i}{\mathcal{A}_u^i}}\, \mathbb{P}_f\left\{Y_i \leq \mathcal{A}_u^i\right\} \leq \frac{\mathcal{A}_0^i}{f(X_i)} \left[\frac{\sqrt{\mathcal{A}_0^i}}{\sqrt{\mathcal{A}_u^i}} \wedge \frac{\sqrt{\mathcal{A}_u^i}}{\sqrt{\mathcal{A}_0^i}}\right] \leq \frac{\mathcal{A}_0^i}{f(X_i)} \left[1 - \frac{|\mathcal{A}_0^i - \mathcal{A}_u^i|}{9M(f)}\right]^{1/2}.$$

The last inequality implies

$$\mathbb{E}_f Z_{n,\theta}^{1/2}(u) \leq \prod_{i:\, X_i \in V_h(y)} \frac{\mathcal{A}_0^i}{f(X_i)} \sqrt{1 - \frac{|f_{u(nh^d)^{-1}}(X_i)|}{9M(f)}}$$

$$\leq e^{\mathcal{N}_h/A(f)} \exp\left\{ -\frac{1}{18M(f)\, nh^d} \sum_{i:\, X_i \in V_h(y)} \big|f_u(X_i)\big| \right\}. \tag{4.7.5}$$

It remains to show

$$\frac{1}{nh^d} \sum_{i:\, X_i \in V_h(y)} \big|f_u(X_i)\big| \geq \lambda_n(h) D_b^{-1} ||u||_1. \tag{4.7.6}$$

Let us remember that $u = (u_p,\, p \in \mathcal{P}_b)$ (where $\mathcal{P}_b$ is defined in (4.1.5)). First, we get from the definition of f_u

$$f_u(x) = u\, K^\top \left(\frac{x-y}{h}\right) = K\left(\frac{x-y}{h}\right) u^\top, \quad \forall x \in [0,1]^d,$$

and, therefore,

$$\frac{1}{nh^d} \sum_{i:\, X_i \in V_h(y)} \big|f_u(X_i)\big| = \frac{1}{nh^d} \sum_{i:\, X_i \in V_h(y)} \left| u\, K^\top \left(\frac{X_i - y}{h}\right) \right|.$$

Assume $u \neq 0$ and put $v = u/||u||_1$. Noting that $|f_v(x)| \leq 1, \quad \forall x \in [0,1]^d$, we have

$$\begin{aligned}
&\frac{1}{nh^d} \sum_{i:\, X_i \in V_h(y)} \big|f_u(X_i)\big| \\
&\geq \frac{1}{nh^d} \sum_{i:\, X_i \in V_h(y)} \left| u\, K^\top \left(\frac{X_i - y}{h}\right) \right| |f_v(X_i)| \\
&= \frac{1}{||u||_1 nh^d} \sum_{i:\, X_i \in V_h(y)} \left| u\, K^\top \left(\frac{X_i - y}{h}\right) K\left(\frac{X_i - y}{h}\right) u^\top \right| \\
&\geq \frac{1}{||u||_1} \left| u\, \frac{1}{nh^d} \sum_{i:\, X_i \in V_h(y)} K^\top \left(\frac{X_i - y}{h}\right) K\left(\frac{X_i - y}{h}\right) u^\top \right|.
\end{aligned}$$

The bound (4.7.6) follows now from Lemma 9. The assertion of the lemma follows from (4.7.5) and (4.7.6).

3. In view of Lemma 8.1, we have

$$\mathbb{E}_f \big| Z_{h,\theta}(u) - Z_{h,\theta}(0) \big| \leq \mathcal{C}_h \|u\|_1, \quad u \in U_n. \tag{4.7.7}$$

Taking into account that $Z_{h,\theta}(0) = 1$ we obtain applying (4.7.7), Fubini's theorem and Markov inequality

$$\begin{aligned}
&\mathbb{P}_f\left\{\int_0^\delta \cdots \int_0^\delta Z_{h,\theta}(v)dv < \frac{1}{2}\delta^{D_b}\right\} \\
&= \mathbb{P}_f\left\{\int_0^\delta \cdots \int_0^\delta \big(Z_{h,\theta}(v) - Z_{h,\theta}(0)\big)dv < -\frac{1}{2}\delta^{D_b}\right\} \\
&\le \mathbb{P}_f\left\{\int_0^\delta \cdots \int_0^\delta \big|Z_{h,\theta}(v) - Z_{h,\theta}(0)\big|dv > \frac{1}{2}\delta^{D_b}\right\} \\
&\le 2\delta^{-D_b}\int_0^\delta \cdots \int_0^\delta \mathbb{E}_f\big|Z_{h,\theta}(v) - Z_{h,\theta}(0)\big|dv \\
&\le 2\mathcal{C}_h\delta
\end{aligned}$$

∎

Proof of Lemma 9 *First step: $\mathcal{M}_{nh}(y)$ is a nonnegative positive matrix.* Let $\mathcal{H}_n, n > 1$ is defined in (4.5.1). First, we prove that

$$\inf_{h\in\mathcal{H}_n} \lambda_n(h) > 0, \quad \forall n > 1. \tag{4.7.8}$$

Suppose that $\exists n_1 > 1, h_{n_1} \in \mathcal{H}_{n_1}$ such that $\lambda_{n_1}\big(h_{n_1}\big) = 0$. Recall that $f_t(x) = tK(h^{-1}(x-y))$ for all $t \in \mathbb{R}^{D_b}$ and note that $\forall \tau \in \mathbb{R}^{D_b}$

$$\begin{aligned}
\tau^\top \mathcal{M}_{n_1 h_{n_1}}(y)\,\tau &= \frac{1}{nh_{n_1}^d}\sum_{i:\, X_i\in V_{h_{n_1}}(y)}\left[\tau\, K^\top\left(\frac{X_i - y}{h_{n_1}}\right)\right]^2 \\
&= \frac{1}{nh_{n_1}^d}\sum_{i:\, X_i\in V_{h_{n_1}}(y)}\big[f_\tau(X_i)\big]^2 \ge 0.
\end{aligned}$$

Since $\lambda_{n_1}\big(h_{n_1}\big)$ is the smallest eigenvalue of the matrix $\mathcal{M}_{n_1 h_{n_1}}(y)$ the assumption $\lambda_{n_1}\big(h_{n_1}\big) = 0$ implies that there exist τ^* belonging to the unit sphere of $\mathbb{R}^{D_b}$ such that

$$\frac{1}{nh_{n_1}^d}\sum_{i:\, X_i\in V_{h_{n_1}}(y)}\big[f_\tau(X_i)\big]^2 = 0.$$

It obviously implies that $f_{\tau^*}(X_i) = 0$ for all $X_i \in V_{h_{n_1}}(y)$. It remains to note that $nh_{n_1}^d \ge \big(b+1\big)^d$ since $h_{n_1} \in \mathcal{H}_n$ and to apply the result obtained in Nemirovski [2000] (page 20). It yields $\tau^* = 0$ and the obtained contradiction proves (4.7.8).

Second step: $\mathcal{M}_{nh}(y) \xrightarrow{n\to\infty} \mathcal{M}$.
Let λ_0 be the smallest eigenvalue of the matrix

$$\mathcal{M} = \int_{[-1/2,1/2]^d} K^\top(x)\, K(x)\, dx$$

whose general term is given by

$$\mathcal{M}_{p,q} = \prod_{j=1}^{d} \int_{-\frac{1}{2}}^{\frac{1}{2}} x_j^{p_j+q_j} dx_j, \quad 0 \le |p|, |q| \le b.$$

Let us prove that

$$\limsup_{n\to\infty} \sup_{h\in\mathcal{H}_n} \left|\lambda_n(h) - \lambda_0\right| = 0. \tag{4.7.9}$$

Put $m = n^{1/d}$ and without loss of generality we will assume that m is integer. Remind that the general term of the matrix $\mathcal{M}_{nh}(y)$ is given by

$$\big(\mathcal{M}_{nh}(y)\big)_{p,q} = \frac{1}{nh^d} \sum_{i:\, X_i\in V_h(y)} \prod_{j=1}^{d} \left(\frac{X_{i_j} - y_j}{h}\right)^{p_j+q_j}.$$

where $X_{i_j} = i_j/m$ for all $j = 1, \dots, d$ and $X_i = \big(X_{i_1}, ..., X_{i_d}\big)$. We get

$$\begin{aligned}
&\frac{1}{nh^d} \sum_{i:\, X_i\in V_h(y)} \prod_{j=1}^{d} \int_{i_j-1}^{i_j} \left(\frac{x_j/m - y_j}{h}\right)^{p_j+q_j} dx_j \\
&\le \frac{1}{nh^d} \sum_{i:\, X_i\in V_h(y)} \prod_{j=1}^{d} \left(\frac{X_{i_j} - y_j}{h}\right)^{p_j+q_j} \\
&\le \frac{1}{nh^d} \sum_{i:\, X_i\in V_h(y)} \prod_{j=1}^{d} \int_{i_j}^{i_j+1} \left(\frac{x_j/m - y_j}{h}\right)^{p_j+q_j} dx_j,
\end{aligned}$$

It yields by change of variables that

$$\begin{aligned}
\prod_{j=1}^{d} \int_{-\frac{1}{2}-2(nh^d)^{-1}}^{\frac{1}{2}} x_j^{p_j+q_j} dx_j \;&\le\; \frac{1}{nh^d} \sum_{i:\, X_i\in V_h(y)} \prod_{j=1}^{d} \left(\frac{X_{i_j} - y_j}{h}\right)^{p_j+q_j} \\
&\le\; \prod_{j=1}^{d} \int_{-\frac{1}{2}}^{\frac{1}{2}+2(nh^d)^{-1}} x_j^{p_j+q_j} dx_j,
\end{aligned} \tag{4.7.10}$$

Note that $nh^d \ge \ln^{\frac{1}{1+d}}(n)$ for any $h \in \mathcal{H}_n$. This together with (4.7.10) yields

$$\limsup_{n\to\infty} \sup_{h\in\mathcal{H}_n} \left|\big(\mathcal{M}_{nh}(y)\big)_{p,q} - \mathcal{M}_{p,q}\right| = 0, \quad 0 \le |p|, |q| \le b.$$

The last result obviously imply (4.7.9).

Third step: Conclusion.
First we show that $\lambda_0 > 0$. Indeed, $\forall \tau \in \mathbb{R}^{D_b}$

$$\tau^\top \mathcal{M} \tau = \int_{[-1/2,1/2]^d} \big[f_\tau(x)\big]^2 dx \geq 0.$$

Since λ_0 is smallest eigenvalue of the matrix $\mathcal{M}$ the assumption $\lambda_0 = 0$ would imply that there exists τ^* belonging to the unit sphere of $\mathbb{R}^{D_b}$ such that $f_{\tau^*} \equiv 0$. Since f_{τ^*} is a polynomial the last identity is possible if and only if $\tau^* = 0$. The obtained contradiction shows that $\lambda_0 > 0$.

Next, note that in view of (4.7.9) there exists n_0 such that $\forall n > n_0$ and $\forall h \in \mathcal{H}_n$, $\lambda_n(h) \geq \lambda_0/2$.

On the other hand in view of (4.7.8) $\min_{n \leq n_0} \inf_{h \in \mathcal{H}_n} \lambda_n(h) > 0$. It remains to define $\lambda > 0$ as

$$\lambda = \min \left(\min_{n \leq n_0} \inf_{h \in \mathcal{H}_n} \lambda_n(h),\ \lambda_0/2 \right).$$

∎

Proof of Lemma 10 Remind that $h_k \leq h_\kappa \leq h^*$ by definition of h_k, h^* and κ (see (4.5.23)). Using Proposition 5 with $h = h_k$, it yields

$$\begin{aligned} \mathbb{E}_f \big|\hat{f}^{(k)}(y) - f(y)\big|^q \mathbb{I}_G &\leq C_q^* \left(\frac{1 \vee Ld\, n h_k^{\beta+d}}{n h_k^d} \right)^q \\ &\leq C_q^* \left(\frac{1 \vee Ld\, n (h^*)^{\beta+d}}{n h_k^d} \right)^q . \end{aligned} \tag{4.7.11}$$

The control of $n(h^*)^{\beta+d}$ requires the following calculation.

$$n(h^*)^{\beta+d} \leq 1 + \frac{b-\beta}{(b+d)(\beta+d)} \ln n = \rho_n(\beta) \tag{4.7.12}$$

where $\rho_n(\beta)$ is the price to pay for adaptation defined in (4.1.10). By definition of h_k, we have

$$\begin{aligned} 1 + \kappa \ln 2 &= 1 + \ln \frac{h_{\max}}{h_k} \geq 1 + \ln \frac{h_{\max}}{h*} \\ &\geq 1 + \frac{b-\beta}{(b+d)(\beta+d)} \ln n - \frac{1}{\beta+d} \ln \big[c\big(1 + (b-\beta)\ln n\big)\big]. \end{aligned}$$

Using the classical inequality $\ln(1+x) \le x$ and $c \le 1/(\beta+d)(b+d)$, we obtain with the last inequality

$$\frac{\beta+d-1}{\beta+d}\rho_n(\beta) \le 1+\kappa\ln 2 \le 1+k\ln 2, \forall k \ge \kappa. \tag{4.7.13}$$

According to (4.7.11),(4.7.12) and (4.7.13), Lemma 10 is proved. ∎

Proof of Lemma 11 Note that for any $k \ge \kappa+1$ and by definition of $\hat{k}$ (4.3.10)

$$\big\{\hat{k}=k\big\} = \cup_{l\ge k}\Big\{\big|\hat{f}^{(k-1)}(y)-\hat{f}^{(l)}(y)\big| > \hat{M}S_n(l)\Big\}.$$

Note that $S_n(l)$ is monotonically increasing in l and, therefore,

$$\begin{aligned}\big\{\hat{k}=k\big\} \;\subseteq\; & \Big\{\big|\hat{f}^{(k-1)}(y)-f(y)\big| > 2^{-1}\hat{M}S_n(k-1)\Big\} \\ & \cup\Big[\cup_{l\ge k}\Big\{\big|\hat{f}^{(l)}(y)-f(y)\big| > 2^{-1}\hat{M}S_n(l)\Big\}\Big].\end{aligned}$$

Taking into account that the event G implies the realization of the event $\hat{M} \ge M(f)/2 \ge A/2$ we come to the following inequality: for any $k \ge \kappa+1$

$$\begin{aligned}\mathbb{P}\big(\hat{k}=k,\,G\big) \;\le\; & \mathbb{P}\Big\{\big|\hat{f}^{(k-1)}(y)-\hat{f}(y)\big| > 4^{-1}M(f)\,S_n(k-1),\,G\Big\} \\ & + \sum_{l\ge k}\mathbb{P}\Big\{\big|\hat{f}^{(l)}(y)-f(y)\big| > 4^{-1}M(f)\,S_n(l),\,G\Big\}. \qquad (4.7.14)\end{aligned}$$

Note that the definition of $S_n(l)$ yields

$$nh_l^d\,S_n(l) \ge 432D_b^3(32qd+16)\lambda^{-1}(h_l)\big[1+\ln\big(h_{\max}/h_l\big)\big].$$

Thus, applying Proposition 4 and Lemma 9 with $J_1 = Ld(1+6D_b^2)/6A(f)D_b^2$ and $\varepsilon = J_1\lambda^{-1}(h_l)\big[1+\ln\big(h_{\max}/h_l\big)\big]$, we obtain $\forall l \ge k-1$

$$\begin{aligned}\mathbb{P}\Big\{\big|\hat{f}^{(l)}(y)-f(y)\big| > (M(f)/4)\,S_n(l),\,G\Big\} \;&\le\; \mathfrak{B}(A,M)\mathcal{E}(h_l)\,[h_{\max}/h_l]^{-8qd-4} \\ &=\; \mathfrak{B}(A,M)\mathcal{E}(h_l)2^{-l(8qd+4)}. \qquad (4.7.15)\end{aligned}$$

Note that $b_{h_l} \le Ldh_l^\beta$ since $f \in \mathbb{H}_d(\beta,L,A,M)$ and, therefore,

$$\mathcal{N}(h_l) \le Ldn(h_l)^{\beta+d} \le Ldn(h_\kappa)^{\beta+d} \le Ldn(h^*)^{\beta+d},\ \ \forall l \ge k-1. \tag{4.7.16}$$

Here we have also used that $k \ge \kappa+1$. We obtain from (4.7.14), (4.7.15) and (4.7.16) that $k \ge \kappa+1$

$$\mathbb{P}\big(\hat{k}=k,\,G\big) \le J_2\mathfrak{B}(A,M)\exp\big\{J_1n(h^*)^{\beta+d}\big\}2^{-(k-1)(8qd+4)},$$

where $J_2 = (1-2^{-(8qd+4)})^{-1}$. ∎

Proof of Lemma 12 Put for any $p \in \mathcal{P}_b$

$$W_{ni}^p(y) = p_1!...p_d! \, \frac{h_{\max}^{d-|p|}}{n} \, K^\top(0) \, \mathcal{M}_{nh_{\max}}^{-1}(y) \, K\left(\frac{X_i - y}{h_{\max}}\right) \mathbb{I}_{V_{\max}(y)}(X_i),$$

and note that $\tilde{\delta}_p = \sum_{i=1}^d 2Y_i \, W_{ni}^p(y)$.

The model (4.1.1) can be rewritten as $2Y_i = f(X_i) + f(X_i)(2U_i - 1)$. Thus, putting $F(X) = \big(f(X_i)\big)_{i=1,\dots,n}$, $V(X) = \big(f(X_i)(2U_i-1)\big)_{i=1,\dots,n}$ and

$$\mathcal{D}(f) = \left(\frac{\partial^{|p|} f(y)}{\partial y_1^{p_1} \cdots \partial y_d^{p_d}}, \, p \in \mathcal{P}_\beta\right),$$

we obtain

$$|\hat{M} - M(f)| \le ||\tilde{\delta} - \mathcal{D}(f)||_1 \le \|\mathcal{V} \, F(X) - \mathcal{D}(f)\|_1 + \|\mathcal{V} \, V(X)\|_1 \, .$$

Here $\mathcal{V}$ is $D_b \times n$-matrix of general term $\mathcal{V}_{pi} = W_{ni}^p(y)$ and $\|.\|_1$ is the ℓ_1-norm. Let us prove that

$$\mathbb{P}_f\left\{|\hat{M} - M(f)| > M(f)/2\right\} \le \exp\left\{-\frac{n^{\frac{b}{b+d}}}{8\vartheta_2^2 D_b^2}\right\}. \tag{4.7.17}$$

In view of the result proved in Härlde, Hart, Marron, and Tsybakov [1992] and Tsybakov [2008] there exist $\vartheta_1, \vartheta_2 > 0$ such that

$$\begin{aligned} \|\mathcal{V} \, F(X) - \mathcal{D}(f)\|_1 &\le \vartheta_1 h_{\max}^{\beta - \lfloor\beta\rfloor}, \\ \sup_{i,x} |W_{ni}^p(y)| &\le \frac{\vartheta_2}{n h_{\max}^d}, \quad p \in \mathcal{P}_\beta. \end{aligned}$$

Remind that $h_{\max} \xrightarrow{n\to\infty} 0$ and, therefore, $\exists n_0$ such that $\vartheta_1 h_{\max}^{\beta-\lfloor\beta\rfloor} \le M(f)/4$ for any $n \ge n_0$. Note that n_0 can be chosen independent on f since $M(f)/4 \ge A/4$. Thus, we get

$$\begin{aligned} &\mathbb{P}_f\left\{|\hat{M} - M(f)| > M(f)/2\right\} \\ &\le \sum_{p\in\mathbb{N}^d:\, 0\le|p|\le\beta} \mathbb{P}_f\left\{\left|\sum_{X_i\in[0,1]^d} f(X_i)(2U_i-1)W_{ni}^p(y)\right| > \frac{M(f)}{4D_b}\right\}. \end{aligned}$$

Noting that $\big|f(X_i)(2U_i-1)W_{ni}^p(y)\big| \le M(f)\dfrac{\vartheta_2}{nh_{\max}^d}$, applying Höeffding inequality Boucheron, Bousquet, and Lugosi [2004] and the last inequality, we obtain

$$\begin{aligned} &\sum_{p\in\mathbb{N}^d:\, 0\le|p|\le\beta} \mathbb{P}_f\left\{\left|\sum_{X_i\in[0,1]^d} f(X_i)(2U_i-1)W_{ni}^p(y)\right| > \frac{M(f)}{4D_b}\right\} \\ &\le D_b \exp\left\{-\frac{nh_{\max}^d}{8\vartheta_2^2 D_b^2}\right\} = D_b \exp\left\{-\frac{n^{\frac{b}{b+d}}}{8\vartheta_2^2 D_b^2}\right\}. \end{aligned} \tag{4.7.18}$$

Therefore (4.7.17) is proved. Since $|f(y) - A(f)| \leq L d h_{\max}^{\beta} \leq A(f)/4$ for $n \geq n_0$ one has

$$\mathbb{P}_f \left\{ |\hat{A} - A(f)| > A(f)/2 \right\} \leq \mathbb{P}_f \left\{ |\hat{M} - M(f)| > A(f)/4 \right\}.$$

Repeating previous calculations we obtain

$$\begin{aligned} \mathbb{P}_f \left\{ |\hat{A} - A(f)| > A(f)/2 \right\} &\leq D_b \exp\left\{ -\frac{\big[A(f)\big]^2 n^{\frac{b}{b+d}}}{16\big[M(f)\big]^2 \vartheta_2^2 D_b^2} \right\} \\ &\leq D_b \exp\left\{ -\frac{A n^{\frac{b}{b+d}}}{16 M \vartheta_2^2 D_b^2} \right\}. \end{aligned} \tag{4.7.19}$$

Since $\mathbb{P}_f\big(G^c\big) \leq \mathbb{P}_f\big(G_{\hat{A}}^c\big) + \mathbb{P}_f\big(G_{\hat{M}}^c\big)$ the assertion of the lemma follows from (4.7.18) and (4.7.19). ∎

Chapter 5

Huber Estimation

Ce chapitre correspond à l'article Chichignoud [2010c]. L'estimateur de Huber est défini pour la première fois dans le cas non-paramétrique. Une caractéristique importante de cet estimateur est sa *robustesse*. En effet, nous démontrons que l'estimateur de Huber estime la fonction de régression dans le modèle de régression additive (défini par (1.2.2)) pour toute densité g_ξ vérifiant les hypothèses 1. Si les paramètres de régularité β, L sont connus, l'estimateur atteint la vitesse $\varphi_{n,2}(\beta)$. Pour l'adaptation, la vitesse atteinte est $\phi_{n,2}(\beta)$. Le *critère de Huber* s'appuye sur l'idée de la médiane (norme ℓ_1) et la moyenne (norme ℓ_2) au voisinage de 0, ce qui permet d'avoir un comportement plus continu que pour le critère ℓ_1. De plus, celui-ci adopte un comportement *robuste*. Nous accordons une importance toute particulière à la proposition 7 qui est un résultat sur les grandes déviations de l'estimation de Huber.

5.1 Introduction

Let statistical experiment is generated by the observation $Z^{(n)} = (X_i, Y_i)_{i=1,\dots n}, n \in \mathbb{N}^*$, where (X_i, Y_i) satisfies the equation

$$Y_i = f(X_i) + \xi_i, \quad i = 1, \dots, n. \tag{5.1.1}$$

Here $f : [0,1]^d \to \mathbb{R}$ is unknown function and we are interested in estimating f at a given point $y \in [0,1]^d$ from observation $Z^{(n)}$.

The random variables (noise) $(\xi_i)_{i\in 1,\dots,n}$ are supposed to be independent and identically distributed of law ξ of unknown density $g_\xi(.)$ with respect to the Lebesgue's measure $\mathbb{R}$.

The design points $(X_i)_{i\in 1,\dots,n}$ are random, independent and uniformly distributed on $[0,1]^d$. The random vector $(X_i)_{i\in 1,\dots,n}$ is independent of $(\xi_i)_{i\in 1,\dots,n}$.

Along the paper, the unknown function f is supposed to be smooth, in particular, it belongs to the Holder ball of functions $\mathbb{H}_d(\beta, L, M)$ (see Definition 16 below). Here $\beta > 0$

is the smoothness of f, M is the upper bound of f and its partial derivatives and $L > 0$ is Lipschitz constant.

The main goal is to estimate f at a given point $y \in [0,1]^d$, for it, we assume following conditions on the unknown density g_ξ of the noise.

Assumptions 2.

1. *g_ξ is symmetric,*
2. *$g_\xi(0) \geq A > 0$,*
3. *g_ξ is continuous in 0.*

Remark that we do not need the existence of the mean for the noise. We suppose in the first assumption that the noise is symmetric. It is reasonable in practice. The other both assumptions imply that it exits an unique theoretical median equals 0 of the noise. i.e. 0 is the unique solution of the equation $\mathbb{P}(\xi < 0) = 1/2$.

Motivation. In this paper, we deal with robust nonparametric estimation, in particular the idea of the median. We search an estimator which is not sensitive to outliers. It is well-known that the median is more robust than average (see Rousseeuw and Leroy [1987] or Huber and Ronchetti [2009]). In imaging, people use more and more of median filters to reconstruct images (Astola, Egiazarian, Foi, and Katkovnik [2010]).

This idea has also used by Arias-Castro and Donoho [2009]. The robustness of the median is studied for estimation of the no-continuous functions, the authors developed a new method so-Called *Two-scale median filte* to improve the minimax rate of convergence.

The asymptotic normality of the median is used by Brown, Cai, and Zhou [2008] to approximate the regression model by the *wavelet sequence data* and use *BlockJS wavelet* (see Stein [1981] et Cai [1999]) for adaptation. But their conditions on the density g_ξ are stronger than Assumptions 2 and only unidimensional functions are studied. Another new point is studied in this paper : the random design of uniform law on a compact of $\mathbb{R}^d$. It is the first time where a estimator, which does not depend on the density g_ξ, is studied from the initial observations.

The principle of the median falls within the framework of M-estimation Van de Geer [2000]. Indeed, it is well-known that the median is the solution to the problem of minimization of absolute values. The main problem is the discontinuity absolutes values of the derivative at 0. In this sense, Huber developed the so-Called *Huber function* Huber [1964, 1981] for the estimation. This theory has been widely taken up by Rousseeuw [1984] in parametric models.

Today, many areas of applications use the *Huber function* in relative GPS positioning Chang and Guo [2005] or imaging Petrus [1999]. In nonparametric estimation, the Huber approach is introduced by Tsybakov [1982a, 1982b, 1983, 1986] and Härdle and Tsybakov [1988, 1992]. Hall and Jones [1990] and Härdle and Tsybakov [1992] used respectively the

Cross Validation and *Plug-in method* for the L_2-risk and pointwise asymptotic normality. Only, Reiss, Rozenholc, and Cuenod [2009] proposed the adaptation for the pointwise risk over locally almost constant functions (i.e. $\beta \leq 1$).

Recently, several approaches of the selection from the family of linear estimators were proposed, see for instance Goldenshluger and Lepski [2009a, 2009b] and Juditsky, Lepski, and Tsybakov [2009] and the references therein. These technologies are used in the construction of data-driven (adaptive) procedures and they are heavily based on the linearity property. As we already mentioned, the *locally Huber estimators* are completely non-linear and in Section 5.3 we propose the selection rule from this family. It requires, in particular, to develop new non-asymptotical exponential inequalities, which may have an independent interest (Proposition 7).

We propose a method of estimation based on the idea of median estimator with the *Huber function* (5.1.6). Actually, we use a locally parametric method, we approximate the function f by a polynomial f_θ on a neighborhood $V_y(h)$ of size h^d and we estimate the coefficients θ by the *Huber estimator* (5.1.7). We study the maximal risk on the isotropic Hölder spaces and we develop an adaptive procedure based on the Lepski's method for the pointwise estimation.

A possible result of this work is the study of estimating anisotropic functions. Indeed, the methods developed by Kerkyacharian, Lepski, and Picard [2001], Klutchnikoff [2005] and Goldenshluger and Lepski [2008,2009a] are based on the linear properties and they do not agree to these estimators. Another perspective is looking for an oracle inequality for the family of Huber estimators indexed by the bandwidth.

Huber estimation and maximal risk. The first part of the paper is devoted to estimate over $\mathbb{H}_d(\beta, L, M)$ estimation, in particular, when the parameters β, L, M and A are supposed to be known *a priori*. We find the *rate of convergence* (5.1.2) on $\mathbb{H}_d(\beta, L, M)$ and propose the estimator with a deterministic choice of $h_n(\beta, L)$ which achieves the rate (5.1.2). We show that for any $\beta \in \mathbb{R}_+^*$, the rate of convergence is bounded from below by the sequence

$$\varphi_n(\beta) = n^{-\frac{\beta}{2\beta+d}}. \tag{5.1.2}$$

We note that (5.1.2) is the rate of linear estimators. But it is not necessarily optimal according to the density of the noise. For example, if ξ is an uniform random variable, it is well-known that the minimax rate is $n^{-\frac{\beta}{\beta+d}}$. In the Gaussian model, we have the same rate $n^{-\frac{\beta}{2\beta+d}}$ which is minimax.

To construct our estimator, we use so-called *local Huber estimation* which consists in the following. Let

$$V_h(y) = \bigotimes_{j=1}^{d} \big[y_j - h/2, y_j + h/2\big],$$

be the neighborhood around y such that $V_h(y) \subseteq [0,1]^d$, where $h \in (0,1)$ be a given scalar. Fix $b > 0$ (without loss of generality we will assume that b is integer) and let

$$D_b = \sum_{m=0}^{b} \binom{m+d-1}{d-1}. \tag{5.1.3}$$

Let $K(z), z \in \mathbb{R}^d$ be the D_b-dimensional vector of polynomials of the following type (the sign $\top$ below means the transposition):

$$K^\top(z) = \left(\prod_{j=1}^{d} z_j^{p_j}, \ \ (p_1, \ldots, p_d) \in \mathbb{N}^d : \ 0 \leq p_1 + \cdots + p_d \leq b \right).$$

For any $t \in \mathbb{R}^{D_b}$ we define the local polynomial

$$f_t(x) \ = \ t^\top K\left(\frac{x-y}{h} \right) \mathbb{I}_{V_h(y)}(x), \quad x \in [0,1]^d, \tag{5.1.4}$$

where $\mathbb{I}$ denotes the indicator function. The local polynomial f_t can be viewed as an approximation of the regression function f inside of the neighborhood V_h. Introduce the following subset of $\mathbb{R}^{D_b}$

$$\Theta\big(M\big) = \left\{ t \in \mathbb{R}^{D_b} : \ \ \|t\|_1 \leq M \right\}, \tag{5.1.5}$$

where $\|.\|_1$ is ℓ_1-norm on $\mathbb{R}^{D_b}$. Consider the *Huber function*,

$$Q(z) = \frac{z^2}{2}\, \mathbb{I}_{|z| \leq 1} + \left(|z| - \frac{1}{2} \right) \mathbb{I}_{|z|>1}, \tag{5.1.6}$$

where $\mathbb{I}$ denotes the indicator function. Put the *Huber criterion*

$$\tilde{m}_h(t) = \tilde{m}_h\big(t, Z^{(n)}\big) := \frac{1}{nh^d} \sum_{i=1}^{n} Q\left(Y_i - f_t(X_i)\right) \mathbb{I}_{\{X_i \in V_h(y)\}} \tag{5.1.7}$$

Let $\breve{\theta}(h)$ be the solution of the following minimization problem :

$$\breve{\theta}(h) = \arg \min_{t \in \Theta(M)} \tilde{m}_h(t). \tag{5.1.8}$$

The *locally Huber estimator* $\breve{f}^h(y)$ of $f(y)$ is defined now as $\breve{f}^h(y) = \breve{\theta}_{0,\ldots,0}(h)$. We note that similar locally parametric approach based on maximum likelihood estimators was recently proposed in Katkovnik and Spokoiny [2008] for *regular statistical models* or based on bayesian estimators was recently proposed in Chichignoud [2010a] for *non-linear models*.

As we see, our construction contains an extra-parameter h to be chosen. To make this choice, we use quite standard arguments. First, we note that in view of $f \in \mathbb{H}_d(\beta, L, M)$

$$\exists \theta = \theta(f, y, h) \in \Theta\big(M\big) : \quad \sup_{x \in V_h(y)} \big|f(x) - f_\theta(x)\big| \le L d h^\beta .$$

Thus, if h is chosen sufficiently small our original model (5.1.1) is well approximated inside of $V_h(y)$ by the "parametric" model

$$\mathcal{Y}_i = f_\theta(X_i) + \xi_i, \quad i = 1, \dots, N_h, \quad N_h = \big(nh^d\big)^{1/2},$$

in which the *Huber estimator* $\breve{\theta}$ is achieves the rate N_h^{-1} (See Theorem 15).

Finally, $h_n(\beta, L)$ is chosen as the solution of the following minimization problem

$$N_h^{-1} + L d h^\beta \to \min_h . \tag{5.1.9}$$

and we show that corresponding estimator $\breve{f}^{h_n(\beta,L)}(y)$ achieves (5.1.2) for $f(y)$ on $\mathbb{H}_d(\beta, L, M)$. For the upper bounds (maximal risk and adaptation), we establish majoration of probability:

$$\mathbb{P}_f \left(\big|\breve{f}^h(y) - f(y)\big| \ge \frac{\varepsilon}{N_h} \right) \le C_1 \exp\left\{ -C_2 \frac{\varepsilon^2}{A + B\varepsilon} \right\}.$$

With Bernstein inequality, the well known point is $B\varepsilon$. For adaptation, ε can be large $(\sqrt{\ln n})$, in general, the term $B = B_n$ is such that $B_n \sqrt{\ln n} \xrightarrow{n\to\infty} 0$.

Adaptive estimation. The second part of the paper is devoted to the adaptive minimax estimation over collection of *isotropic* functional classes in the model (5.1.1). To our knowledge, the problem of adaptive estimation, with Huber estimator in pointwise estimation, is not study in the literature. As we mentioned above we consider only the case $\beta \in (0, b]$ and $L \in [l_*, l^*]$ will be studied.

Well-known disadvantage of maximal approach is the dependence of the estimator on the parameters describing functional class on which the maximal risk is determined. In particular, $h_n(\beta, L)$ optimally chosen in view of (5.1.9) depends explicitly on β and L. To overcome this drawback, the maximal adaptive approach was proposed by Lepski [1990, 1991] and Lepski, Mammen, and Spokoiny [1997]. The first question arising in the adaptation (reduced to the problem at hand) can be formulated as follows.

Does there exist an estimator which would be minimax on $\mathbb{H}(\beta, L, M)$ simultaneously for all values of β and L belonging to some given set $\mathfrak{B} \subseteq \big[\mathbb{R}_+ \setminus 0\big] \times \big[\mathbb{R}_+ \setminus 0\big]$?

We can show that the answer of this question is negative, that is typical for the estimation of the function at a given point Lepski and Spokoiny [1997]. This answer can be reformulated in the following manner: the family of rates of convergence $\big\{\varphi_n(\beta, L),\ (\beta, L) \in \mathfrak{B}\big\}$ is unattainable for the problem under consideration.

Thus, we need to find another family of normalizations for maximal risk which would be attainable and, moreover, optimal in view of some criterion of optimality. Nowadays, the most developed criterion of optimality is due to Klutchnikoff [2005] in the Gaussian model.

The most important step is to find an estimator, called *adaptive*, which attains this family of normalizations:

$$\phi_n(\beta, L) = \left(\frac{\rho_n(\beta, L)}{n}\right)^{\frac{\beta}{2\beta+d}} \tag{5.1.10}$$

with $\rho_n(\beta, L) = \left(1 + \frac{2(b-\beta)}{(2\beta+d)(2b+d)} \ln n\right)^{1/2}$. The factor ρ_n can be considered as *price to pay for adaptation* Lepski [1990].

In the present paper, we construct such estimator using general adaptation scheme due to Lepski, Mammen, and Spokoiny [1997] (so-called *Lepski's method*). To our knowledge, it is the first time that this method is applied with *Huber estimator* and in the statistical model with the unknown density of the noise. However, the limitation concerning the consideration of isotropic classes of functions is also due to the use of Lepski's procedure. It seems that to be able to treat the adaptation over the scale of anisotropic classes, another scheme should be applied Kerkyacharian, Lepski, and Picard [2001], Klutchnikoff [2005] and Goldenshluger and Lepski [2008]. For the model (5.1.1) this problem is still open.

This paper is organized as follows. In Section 5.2 we present the results concerning maximal risk and Section 5.3 is devoted to the adaptive estimation. The proofs of main results are given in Section 5.4 (upper bound), technical lemmas are postponed to Appendix.

5.2 Maximal Risk on $\mathbb{H}_d(\beta, L, M)$

In this section, we present several results concerning Maximal risk. We propose the estimator which is bounded the maximal risk on this class of functions under some additional restrictions imposed on these parameters. For any $(p_1, ..., p_d) \in \mathbb{N}^d$ we denote $\vec{p} = (p_1, ..., p_d)$ and $|\vec{p}| = p_1 + ... + p_d$.

Definition 16. *Fix $\beta > 0$, $L > 0$, $M > 0$ and let $\lfloor \beta \rfloor$ be the largest integer strictly less than β. The* isotropic Hölder class $\mathbb{H}_d(\beta, L, M)$ *is the set of functions $f : [0,1]^d \to \mathbb{R}$ having on $[0,1]^d$ all partial derivatives of order $\lfloor \beta \rfloor$ and such that $\forall x, y \in [0,1]^d$*

$$\begin{aligned}
\sum_{m=0}^{\lfloor \beta \rfloor} \sum_{|\vec{p}|=m} \sup_{x \in [0,1]^d} \left| \frac{\partial^{|\vec{p}|} f(x)}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} \right| &\leq M, \\
\left| \frac{\partial^{|\vec{p}|} f(x)}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} - \frac{\partial^{|\vec{p}|} f(y)}{\partial y_1^{p_1} \cdots \partial y_d^{p_d}} \right| &\leq L \left[\|x - y\|_1 \right]^{\beta - \lfloor \beta \rfloor}, \quad \forall \, |\vec{p}| = \lfloor \beta \rfloor.
\end{aligned}$$

To measure the performance of estimation procedures on $\mathbb{H}_d(\beta, L)$ we will use maximal approach.

Let $\mathbb{E}_f = \mathbb{E}_f^n$ be the mathematical expectation with respect to the probability law of the observation $Z^{(n)}$ satisfying (5.1.1). Firstly, we define the maximal risk on $\mathbb{H}_d(\beta, L)$ corresponding to the estimation of the function f at a given point $y \in [0,1]^d$.

Let $\tilde{f}_n$ be an arbitrary estimator built from the observation $Z^{(n)}$. Set $\forall r > 0$

$$R_{n,r}\big[\tilde{f}_n, \mathbb{H}_d(\beta, L)\big] = \sup_{f \in \mathbb{H}_d(\beta,L)} \mathbb{E}_f \big|\tilde{f}_n(y) - f(y)\big|^r.$$

The quantity $R_{n,r}\big[\tilde{f}_n, \mathbb{H}_d(\beta, L)\big]$ is called *maximal risk* of the estimator $\tilde{f}_n$ on $\mathbb{H}_d(\beta, L)$.

Upper bound for maximal risk. Let us start with the construction of our estimator based on *local Huber estimation*.

Let

$$\bar{h} = (L^2 n)^{-\frac{1}{2\beta+d}}, \ j = 1, \dots, d. \tag{5.2.1}$$

The next theorem shows how to construct the estimator basing on locally Huber approach which achieves the rate (5.1.2). Let $\bar{f}^{\bar{h}}(y) = \breve{\theta}_{0,\dots,0}\big(\bar{h}\big)$ is given by (5.1.5), (5.1.7) and (5.1.8) with $h = \bar{h}$.

Theorem 15. *Let $\beta > 0$, $L > 0$ and $M > 0$ be fixed. Then for any density g_ξ verified Assumptions 2, we have*

$$\limsup_{n\to\infty} \varphi_n^{-r}(\beta) R_{n,r}\Big[\bar{f}^{\bar{h}}(y), \mathbb{H}_d(\beta, L, M)\Big] < \infty, \quad \forall r \geq 1,$$

Remark 15. *We can see that, asymptotically, the rate of convergence of risk is $\varphi_n(\beta)$. We can show in some models that $n^{-\frac{\beta}{2\beta+d}}$ is minimax (gaussian and Cauchy models, Tsybakov [2008]). Tsybakov [1982a] showed lower bounds for minimax probability risk over Hölder classes with the additive model (5.1.1). But if g_ξ is the density of uniform random variable, then the minimax rate is $n^{-\frac{\beta}{\beta+d}}$, so our Huber estimator is not optimal in this sense. The advantage of this approach is that we do not assume knowledge of the density g_ξ, only a few conditions (see Assumptions 2), we talk "robust estimation".*

5.3 Bandwidth Selector of Huber Estimator

This section is devoted to the adaptive estimation over the collection of the classes $\Big\{\mathbb{H}_d(\beta, L, M)\Big\}_{\beta,L}$, we suppose M known. We will not impose any restriction on possible

value of L, but we will assume that $\beta \in (0, b]$, where b, as previously, is an arbitrary a priori chosen integer.

We start with formulating the remark showing that there is not optimally adaptive estimator (here we follow the terminology introduced in Lepski [1990, 1992a, 1992b]). It means that there is not an estimator which would be minimax simultaneously for several values of parameter β even if L is supposed to be fixed. This result does not require any restriction on β as well. Let $\Psi = \{\psi_n(\beta)\}_{\beta\in(0,b]}$ be a given family of normalizations.

Definition 17. *The family Ψ is called admissible if there exists an estimator $\breve{f}_n$ such that for some $L, M > 0$*

$$\limsup_{n\to\infty} \psi_n^{-r}(\beta)\, R_{n,r}\big(\breve{f}_n, \mathbb{H}_d(\beta, L, M)\big) < \infty, \quad \forall \beta \in (0, b]. \tag{5.3.1}$$

The estimator $\breve{f}_n$ satisfying (5.3.1) is called Ψ-attainable. The estimator $\breve{f}_n$ is called Ψ-adaptive if (5.3.1) holds for any $L > 0$ and $M > 0$.

Remark 16. *Note that the family of rates of convergence $\{\varphi_n(\beta)\}_{\beta\in(0,b]}$ is not admissible. We can show this assertion with Lepski and Spokoiny [1997] in Gaussian model.*

Let Φ be the following family of normalizations:

$$\phi_n(\beta) = \left(\frac{\rho_n(\beta)}{n}\right)^{\frac{\beta}{2\beta+d}}, \quad \rho_n(\beta) = \left(1 + \frac{2(b-\beta)}{(2\beta+d)(2b+d)} \ln n\right)^{1/2}, \quad \beta \in (0, b].$$

We remark that $\phi_n(b) = \varphi_n(b)$ and $\rho_n(\beta) \sim \sqrt{\ln n}$ for any $\beta \neq b$.

Remark 17. *We can show that this family Φ is adaptive optimal using Klutchnikoff's Criterion Klutchnikoff [2005] in Gaussian model.*

Construction of Φ-adaptive estimator. As it was already mentioned in Introduction, the construction of our estimation procedure consists of several steps. First, we determine family of *locally Huber estimators*. Next, based on so-called *Lepski's method*, we propose data-driven selection from this family.

We take $\breve{f}^h$ the estimator given by (5.1.5), (5.1.7) and (5.1.8), so the family of locally bayesian estimator $\breve{\mathcal{F}}$ is defined now as follows. Put $h_{\max} = n^{-\frac{1}{2b+d}}$ and

$$h_k = 2^{-k} h_{\max}, \quad k = 0, \dots, \mathbf{k}_n,$$

where $\mathbf{k}_n$ is the largest integer such that $h_{\mathbf{k}_n} \geq h_{\min} = n^{-1/d} \ln^{\frac{b}{d(2b+d)}} n$. Set

$$\breve{\mathcal{F}} = \left\{ \breve{f}^{(k)}(y) = \breve{\theta}_{0,\dots,0}(h_k), \quad k = 0, \dots, \mathbf{k}_n \right\}. \tag{5.3.2}$$

We put $\breve{f}^*(y) = \breve{f}^{(\hat{k})}(y)$, where $\breve{f}^{(\hat{k})}(y)$ is selected from $\breve{\mathcal{F}}$ in accordance with the rule:

$$\hat{k} = \inf\left\{k = \overline{0, \mathbf{k}_n} : \ \left|\breve{f}^{(k)}(y) - \breve{f}^{(l)}(y)\right| \le CS_n(l), \ \ l = \overline{k+1, \mathbf{k}_n}\right\}. \tag{5.3.3}$$

Here we have used the following notations.

$$C = 8rd\,\frac{D_b^2}{\lambda}\left(8 + \frac{4\sqrt{\lambda}}{3D_b}\right), \quad S_n(l) = \left[\frac{1 + l\ln 2}{n\big(h_l\big)^d}\right]^{1/2}, \quad l = 0, 1, \ldots, \mathbf{k}_n, \tag{5.3.4}$$

where $\lambda > 0$ is a constant defined in Lemma 13 and D_b defined in (5.1.3).

Theorem 16. *Let $b > 0, M > 0$ be fixed. Then, for any density g_ξ verified Assumptions 2, for any $\beta \in (0, b]$, $L > 0$ and $r \ge 1$*

$$\limsup_{n\to\infty} \phi_n^{-r}(\beta) R_{n,r}\Big[\breve{f}^*(y), \mathbb{H}_d(\beta, L, M)\Big] < \infty.$$

Remark 18. *The assertion of the theorem means that the proposed estimator $\breve{f}^*(y)$ is Φ-adaptive. It implies in particular that the family of normalizations Φ is admissible.*

Remark 19. *Remark that M is involved in construction of the set of coefficients $\Theta(M)$, for example we can replace M by $\ln n$. The constant λ depends on $g_\xi(0)$ and $\|g_\xi\|_\infty$ (see Proof of Lemma 13), unknown in practice, but we can calibrate the constant C.*

5.4 Proofs of Main Results: Upper Bounds

Let $\mathcal{H}_n, n > 1$ be the following subinterval of $(0, 1)$.

$$\mathcal{H}_n = \left[\frac{\big(\ln n\big)^{\frac{1}{d}}}{n^{1/d}}, \ \left(\frac{1}{\ln n}\right)^{\frac{1}{2b+d}}\right]. \tag{5.4.1}$$

Later on, we will consider only the values of h belonging to $\mathcal{H}_n$. Put also $N_h = \big(nh^d\big)^{1/2}$.

5.4.1 Auxiliary Results: Large Deviations for M-estimators

Let $h \in \mathcal{H}_n$, put $\theta = \theta(f, y, h) = \big\{\theta_{\vec{p}} : \ \vec{p} \in \mathbb{N}^d : \ 0 \le |\vec{p}| \le b\big\}$, where $\theta_0 = \theta_{0,\ldots,0} = f(y)$ and

$$\theta_{\vec{p}} = \frac{\partial^{|\vec{p}|} f(y)}{\partial y_1^{p_1} \cdots \partial y_d^{p_d}} \frac{h^{|\vec{p}|}}{p_1! \ldots p_d!}, \ p \in \mathbb{N}^d : \ 0 < |\vec{p}| \le b.$$

Remind the agreement which we follow in the present paper: if the function f and vector $\vec{p}$ be such that $\partial^{|\vec{p}|} f$ does not exist we put $\theta_{\vec{p}} = 0$.

Let $f_\theta(x)$, given by (5.1.4), be the local polynomial approximation of f inside $V_h(y)$ and let b_h be the corresponding approximation error, i.e.

$$b_h := \sup_{x\in V_h(y)} \big|f_\theta(x) - f(x)\big|.$$

Introduce for any function f

$$M(f) = \sum_{m=0}^{b} \sum_{p_1+\ldots+p_d=m} \left| \frac{\partial^m f(y)}{\partial x_1^{p_1}\cdots\partial x_d^{p_d}} \right|. \tag{5.4.2}$$

The following agreement will be used in the sequel: if the function f and $m \geq 1$ be such that $\partial^m f$ does not exist, we will put formally $\partial^m f = 0$ in the definition of $M(f)$.

Put

$$q(z) := Q'(z) = \mathbb{I}_{]1,+\infty[}(z) - \mathbb{I}_{]-\infty,-1[}(z) + z\,\mathbb{I}_{[-1,1]}(z), \quad z\in\mathbb{R}, \tag{5.4.3}$$

where Q' is the derivative of *Huber function* Q defined in (5.1.6). Set $\mathcal{S}_b = \big\{\vec{q}\in\mathbb{N}^d : \ |\vec{q}|\leq b\big\}$ and for any $t\in\Theta(M)$ define the so-called *partial derivative of Huber criterion*

$$\tilde{D}_h^{\vec{p}}(t) = \frac{\partial^{|\vec{p}|}\tilde{m}_h(t)}{\partial t_{\vec{p}}} = -\frac{1}{nh^d}\sum_{i=1}^{n} q\big(Y_i - f_t(X_i)\big)\left(\frac{X_i - y}{h}\right)^{\vec{p}} \mathbb{I}_{X_i\in V_y(h)}, \quad \vec{p}\in\mathcal{S}_b, \tag{5.4.4}$$

where $x^{\vec{p}} = x_1^{p_1}\ldots x_d^{p_d}$, $\forall x\in[0,1]^d$. Put the *expectation of partial derivative of Huber criterion*:

$$\begin{aligned}
\mathcal{E}_h^{\vec{p}}(t) &:= \mathbb{E}_f\Big[\tilde{D}_h^{\vec{p}}(t)\Big] \\
&= -\frac{1}{h^d}\int_{V_y(h)}\left(\frac{x-y}{h}\right)^{\vec{p}}\left[\int_{1+f_t(x)-f(x)}^{+\infty} g_\xi(z)dz - \int_{-\infty}^{-1+f_t(x)-f(x)} g_\xi(z)dz\right. \\
&\quad \left. + \int_{-1+f_t(x)-f(x)}^{1+f_t(x)-f(x)} \big(z - f_t(x) + f(x)\big) g_\xi(z)dz\right] dx, \quad \vec{p}\in\mathcal{S}_b,
\end{aligned} \tag{5.4.5}$$

We decompose $\mathcal{E}_h^{\vec{p}}$ in two parts: $\mathcal{E}_h^{\vec{p}}(t) = D_h^{\vec{p}}(t) + \Delta_h^{\vec{p}}(t)$, where

$$\begin{aligned}
D_h^{\vec{p}}(t) &= -\frac{1}{h^d}\int_{V_y(h)}\left(\frac{x-y}{h}\right)^{\vec{p}}\left[\int_{1}^{+\infty} g_\xi\big(z + f_{t-\theta}(x)\big)dz - \int_{-\infty}^{-1} g_\xi\big(z + f_{t-\theta}(x)\big)dz\right. \\
&\quad \left. + \int_{-1}^{1} z\, g_\xi\big(z + f_{t-\theta}(x)\big)dz\right] dx, \quad \vec{p}\in\mathcal{S}_b.
\end{aligned} \tag{5.4.6}$$

and to simplify we note

$$\Delta_h^{\vec{p}}(t) = \mathcal{E}_h^{\vec{p}}(t) - D_h^{\vec{p}}(t), \quad \vec{p}\in\mathcal{S}_b, \tag{5.4.7}$$

$\Delta_h(t) = \big(\Delta_h^{\vec{p}}(t)\big)^\top_{\vec{p}\in\mathcal{S}_b}$ and $D_h(t) = \big(D_h^{\vec{p}}(t)\big)^\top_{\vec{p}\in\mathcal{S}_b}$ are respectively so-called *bias term* and *deterministic criterion*. Finally, let us define the *partial derivative of deterministic criterion* by the following notation

$$\nabla_h^{\vec{p},\vec{q}}(t) = \frac{\partial^{|\vec{q}|} D_h^{\vec{p}}(t)}{\partial t_{\vec{q}}} = \frac{1}{h^d}\int_{V_h(y)} \left(\frac{x-y}{h}\right)^{\vec{p}+\vec{q}} \int_{-1}^{1} g_\xi\big(z + f_{t-\theta}(x)\big)dz\,dx, \quad \vec{p},\vec{q}\in\mathcal{S}_b^2. \tag{5.4.8}$$

We call *Jacobian matrix of deterministic criterion*

$$J_D(t) = \Big(\nabla_h^{\vec{p},\vec{q}}(t)\Big)_{\vec{p},\vec{q}\in\mathcal{S}_b^2}. \tag{5.4.9}$$

The main property of the matrix $J_D(.)$ is that it exists a neighborhood $\mathcal{V}$ of θ does not depend on n such that $J_D(.)$ is invertible on $\mathcal{V}$ (see Lemma 13). The next proposition is the milestone for all results proved in the paper.

Put the random sets for any $h, z > 0$

$$G_z^h = \big\{\breve{\theta}(h) \in \mathcal{B}\big(\theta, z\big)\big\}, \quad \mathcal{B}\big(\theta, z\big) = \big\{t \in \theta(M) : \ \|t-\theta\|_2 \le z\big\}. \tag{5.4.10}$$

Proposition 7. *For any $n \in \mathbb{N}^*$, $h \in \mathcal{H}_n$, f such that $M(f) < \infty$ and $\delta > 0$ such that $\mathcal{B}\big(\theta,\delta\big) \subseteq \mathcal{V}$, then $\forall \varepsilon \ge D_b\,\varepsilon_0(h)/\sqrt{\lambda}$ we have*

$$\mathbb{P}_f\left(N_h\big|\breve{f}^h(y) - f(y)\big| \ge \varepsilon,\ G_\delta^h\right) \le \Sigma \exp\left\{-\frac{\Big(\varepsilon\sqrt{\lambda}D_b^{-1} - \varepsilon_0(h)\Big)^2}{8 + \frac{8}{3N_h}\varepsilon\sqrt{\lambda}D_b^{-1}}\right\},$$

where $\varepsilon_0(h) = \big(1 \vee b_h\,N_h\big)$, $\breve{f}^h(y)$ defined in (5.1.5), (5.1.7), (5.1.8), λ is a positive constant defined in Lemma 13 and Σ is defined in Lemma 15.

The next proposition provides us with upper bound for the risk of a locally Huber estimator. Put

$$\bar{C}_r = \frac{D_b^r}{\lambda^{r/2}}\left[1 + \int_0^{+\infty}(D_b^{-1}\sqrt{\lambda}\eta + 1)^{r-1}\exp\left\{-\frac{\eta^2}{8\left(1+\eta+L\sqrt{\lambda}D_b^{-1}\right)}\right\}d\eta\right], \quad r \ge 1. \tag{5.4.11}$$

Proposition 8. *For any $n \in \mathbb{N}^*$, $h \in \mathcal{H}_n$, $\delta > 0$ such that $\mathcal{B}\big(\theta,\delta\big) \subseteq \mathcal{V}$ and f such that $M(f) < \infty$, then*

$$\mathbb{E}_f\big|\breve{f}^h(y) - f(y)\big|^r \mathbb{I}_{G_\delta^h} \le \bar{C}_r\,\varepsilon_0^r(h)\,N_h^{-r}, \quad r \ge 1.$$

The proof of this proposition is elementary by integration of Proposition 7.

5.4.2 Proof of Proposition 7

Auxiliary lemmas. We give the following lemma concerning *deterministic criterion*.

Lemma 13. *For any g_ξ verified Assumptions 2, then*

1. *$J_D(\theta)$ is positive definite matrix.*
2. *θ is unique solution of $D_h(.) = (0, ..., 0)$.*
3. *$\forall h \in \mathcal{H}_n$, it exists a neighborhood $\mathcal{V}$ of θ, which does not depend on n, such that*

$$\|\tilde{\theta} - \theta\|_2 \leq \lambda^{-1/2} \left\| D_h\big(\tilde{\theta}\big) - D_h(\theta) \right\|_2, \quad \tilde{\theta} \in \mathcal{V},$$

where $\lambda = \inf_{\tilde{\theta} \in \mathcal{V}} \lambda_{\tilde{\theta}}$, $\lambda_{\tilde{\theta}}$ is the smallest eigenvalue of matrix $J_D\big(\tilde{\theta}\big)^\top J_D\big(\tilde{\theta}\big)$ and $\|.\|_2$ is the ℓ_2-norm on $\mathbb{R}^{D_b}$.

The proof is given in Appendix. Let us give a lemma which allows to control the *bias term*.

Lemma 14. *For any $h \in \mathcal{H}_n$ and f such that $M(f) < \infty$, we have*

$$\left|\Delta_h^{\vec{p}}(t)\right| \leq b_h, \quad \vec{p} \in \mathcal{S}_b,\ t \in \theta_M.,$$

where $\|.\|_\infty$ is the sup-norm

The next result allows to control the large deviations of *partial derivatives of Huber criterion*.

Lemma 15. *For any f such that $M(f) < \infty$ and $h \in \mathcal{H}_n$, we have $\forall z \geq \big(1 \vee b_h\, N_h\big)$*

$$\mathbb{P}_f\left(N_h \sup_{t \in \Theta(M)} \left| \tilde{D}_h^{\vec{p}}(t) - D_h^{\vec{p}}(t) \right| \geq z \right) \leq \Sigma \exp\left\{ -\frac{(z - b_h\, N_h)^2}{8 + \frac{8}{3N_h} z} \right\}, \quad \vec{p} \in \mathcal{S}_b,$$

where Σ is defined in the proof.

The proof (given in Appendix) is based on a chaining argument and Bernstein inequality. It is required that the *Huber criterion* is C^1.

Lemma 16. *For any $f \in \mathbb{H}_d(\beta, L, M)$, $\delta > 0$ such that $\mathcal{B}(\theta, \delta) \subseteq \mathcal{V}$ and $n \in \mathbb{N}^*$, $h \in \mathcal{H}_n$ such that*

$$b_h \leq \frac{\varkappa_\delta}{2\sqrt{D_b}}, \qquad N_h \geq \frac{2\sqrt{D_b}}{\varkappa_\delta}, \qquad \varkappa_\delta = \inf_{t \in \Theta(M) \setminus \mathcal{B}(\theta, \delta)} \frac{\|D_h(t)\|_2}{2},$$

then

$$\mathbb{P}_f\left[\big(G_\delta^h\big)^c \right] \leq D_b \Sigma \, \exp\left\{ -\frac{n h^d \varkappa_\delta^2}{4 D_b \left(8 + 4\varkappa_\delta / (3\sqrt{D_b})\right)} \right\},$$

where $\big(G_\delta^h\big)^c$ is the additional event of G_δ^h, $\mathcal{V}$ defined in Lemma 13.

Proof of Proposition 7. Remark that $\tilde{D}_h\big(\breve{\theta}\big) = (0, ..., 0)$ where $\tilde{D}_h(.) = \big(\tilde{D}_h^{\vec{p}}(.)\big)_{\vec{p}\in\mathcal{S}_b}^{\top}$ defined in (5.4.4). The main idea of this proof is that if $\tilde{D}_h(t) \xrightarrow[n\to\infty]{N_h^{-1}} \mathcal{E}_h(t)$ in probability uniformly in t then we have $\breve{\theta}(h) \xrightarrow[n\to\infty]{N_h^{-1}} \theta$. This is the idea of M-estimation Van de Geer [2000]. The definition of $\breve{\theta}(h)$ and $\theta = \theta(f, y, h)$ implies that $\forall \varepsilon \geq D_b\sqrt{\lambda}\,\varepsilon_0(h)$

$$\begin{aligned}\mathbb{P}_f\left(N_h\big|\breve{f}^h(y) - f(y)\big| \geq \varepsilon,\, G_\delta^h\right) &\leq \mathbb{P}_f\left(N_h\big|\breve{\theta}_0(h) - \theta_0\big| \geq \varepsilon,\, G_\delta^h\right)\\ &\leq \mathbb{P}_f\left(N_h\sqrt{D_b}\big\|\breve{\theta}(h) - \theta\big\|_2 \geq \varepsilon,\, G_\delta^h\right).\end{aligned}$$

Under the event G_δ^h we have $\breve{\theta}(h) \in \mathcal{V}$, according Lemma 13 we obtain that

$$\mathbb{P}_f\left(N_h\big|\breve{f}^h(y) - f(y)\big| \geq \varepsilon,\, G_\delta^h\right) \leq \mathbb{P}_f\left(N_h\frac{\sqrt{D_b}}{\sqrt{\lambda}}\Big\|D_h\big(\breve{\theta}(h)\big) - D_h(\theta)\Big\|_2 \geq \varepsilon\right).$$

Remember that $D_h(\theta) = 0$ and $\tilde{D}_h\big(\breve{\theta}(h)\big) = 0$. Using the last inequality, we get

$$\begin{aligned}\mathbb{P}_f\left(N_h\big|\breve{f}^h(y) - f(y)\big| \geq \varepsilon,\, G_\delta^h\right) &\leq \mathbb{P}_f\left(N_h\frac{\sqrt{D_b}}{\sqrt{\lambda}}\Big\|D_h\big(\breve{\theta}(h)\big) - \tilde{D}_h\big(\breve{\theta}(h)\big)\Big\|_2 \geq \varepsilon\right)\\ &\leq \sum_{\vec{p}\in\mathcal{S}_b}\mathbb{P}_f\left(N_h\sup_{t\in\Theta(M)}\Big|\tilde{D}_h^{\vec{p}}(t) - D_h^{\vec{p}}(t)\Big| \geq \frac{\varepsilon\sqrt{\lambda}}{D_b}\right).\end{aligned}$$

Applying Lemma 15 with $z = \varepsilon\sqrt{\lambda}/D_b$ and last inequality, finally we have the assertion of Proposition 7

$$\mathbb{P}_f\left(N_h\big|\breve{f}^h(y) - f(y)\big| \geq \varepsilon,\, G_\delta^h\right) \leq \Sigma\exp\left\{-\frac{\Big(\varepsilon\sqrt{\lambda}D_b^{-1} - \varepsilon_0(h)\Big)^2}{8 + \frac{8}{3N_h}\varepsilon\sqrt{\lambda}D_b^{-1}}\right\}.$$

■

5.4.3 Proof of Theorem 15

By definition of $\bar{h} = (L^2n)^{-\frac{1}{2\beta+d}}$, we have $b_{\bar{h}}\,N_{\bar{h}} = d$, $\quad N_{\bar{h}}^{-r} = L^{\frac{rd}{2\beta+d}}\varphi_n^r(\beta)$. So we obtain that the term $\varepsilon_0\big(\bar{h}\big)$ defined in Proposition 7 can be written as $\varepsilon_0\big(\bar{h}\big) = d$. We get

$$\mathbb{E}_f\big|\bar{f}^{\bar{h}}(y) - f(y)\big|^r = \mathbb{E}_f\big|\bar{f}^{\bar{h}}(y) - f(y)\big|^r\mathbb{I}_{G_\delta^{\bar{h}}} + \mathbb{E}_f\big|\bar{f}^{\bar{h}}(y) - f(y)\big|^r\mathbb{I}_{\left(G_\delta^{\bar{h}}\right)^c}. \tag{5.4.12}$$

The right hand side is controlled by Lemma 16. Indeed, we can use the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}_f\big|\bar{f}^{\bar{h}}(y)-f(y)\big|^r\mathbb{I}_{\left(G_\delta^{\bar{h}}\right)^c} &\le \Big(\mathbb{E}_f\big|\bar{f}^{\bar{h}}(y)-f(y)\big|^{2r}\mathbb{P}_f\Big\{\big(G_\delta^{\bar{h}}\big)^c\Big\}\Big)^{1/2}\\ &\le (2M)^r D_b\Sigma\,\exp\left\{-\frac{nh^d\varkappa_\delta^2}{4D_b\left(8+4\varkappa_\delta/(3\sqrt{D_b})\right)}\right\}. \qquad (5.4.13)\end{aligned}$$

Using Proposition 8, (5.4.12) and (5.4.13), we obtain

$$\mathbb{E}_f\big|\bar{f}^{\bar{h}}(y)-f(y)\big|^r\mathbb{I}_{G_\delta^{\bar{h}}} \le \bar{C}_r d\,L^{\frac{rd}{2\beta+d}}\varphi_n^r(\beta)+(2M)^r D_b\Sigma\,\exp\left\{-\frac{nh^d\varkappa_\delta^2}{4D_b\left(8+4\varkappa_\delta/(3\sqrt{D_b})\right)}\right\}.$$

By passing to the limit in n, the theorem 15 is proved. ∎

5.4.4 Proof of Theorem 16

We start the proof with formulating some auxiliary results whose proofs are given in Appendix. Define $J_1=\sqrt{\lambda}/D_b$ and

$$h^*=\left[\frac{c\,\rho_n^2(\beta)}{L^2d^2\,n}\right]^{\frac{1}{2\beta+d}},\quad c<(2J_1)^{-2}.$$

and let the integer κ be defined as follows.

$$2^{-\kappa}h_{\max}\le h^*<2^{-\kappa+1}h_{\max}. \qquad (5.4.14)$$

Lemma 17. *For any $f\in\mathbb{H}_d(\beta,L,M,A)$ and any $k\ge\kappa+1$*

$$\mathbb{P}_f\big(\hat{k}=k,G_\delta^{h_k}\big)\le J_2\,2^{-2(k-1)rd},$$

where $J_2=D_b\Sigma\big(1+(1-2^{-2rd})^{-1}\big)$.

Proof of Theorem 16. This proof is based on the scheme due to Lepski, Mammen, and Spokoiny [1997]. The definition of h^* and κ (5.4.14) implies that

$$\varepsilon_0\big(h_k\big)\le D_b\frac{\sqrt{1+k\ln 2}}{\sqrt{\lambda}},\quad\forall k\ge\kappa,$$

where $\varepsilon_0(.)$ defined in Propostion 7. Last inequality yields

$$\mathbb{E}_f\big|\check{f}^{(k)}(y)-f(y)\big|^r\mathbb{I}_{G_\delta^{h_k}}\le\bar{C}_rD_b^r\lambda^{-r/2}S_n^r(k),\quad\forall k\ge\kappa. \qquad (5.4.15)$$

To get this result we have applied Proposition 8 with $h = h_k$. We also have

$$\begin{aligned}
&\mathbb{E}_f\big|\breve{f}^{(\hat{k})}(y) - f(y)\big|^r \\
&\quad = \mathbb{E}_f\big|\breve{f}^{(\hat{k})}(y) - f(y)\big|^r \mathbb{I}_{\hat{k}\le\kappa, G_\delta^{h_{\hat{k}}}} + \mathbb{E}_f\big|\breve{f}^{(\hat{k})}(y) - f(y)\big|^r \mathbb{I}_{\hat{k}>\kappa, G_\delta^{h_{\hat{k}}}} \\
&\qquad + \mathbb{E}_f\big|\breve{f}^{(\hat{k})}(y) - f(y)\big|^r \mathbb{I}_{\left(G_\delta^{h_{\hat{k}}}\right)^c} \\
&\quad := R_1(f) + R_2(f) + R_3(f). \qquad (5.4.16)
\end{aligned}$$

First we control R_1. Obviously

$$\big|\breve{f}^{(\hat{k})}(y) - f(y)\big|^r \le 2^{r-1}\big|\breve{f}^{(\hat{k})}(y) - \breve{f}^{(\kappa)}(y)\big|^r + 2^{r-1}\big|\breve{f}^{(\kappa)}(y) - f(y)\big|^r.$$

The definition of $\hat{k}$ yields

$$\big|\breve{f}^{(\hat{k})}(y) - \breve{f}^{(\kappa)}(y)\big|^r \mathbb{I}_{\hat{k}\le\kappa, G_\delta^{h_{\hat{k}}}} \le C^r S_n^r(\kappa),$$

In view of (5.4.15) and definitions of h_κ and h^*, we also get

$$\mathbb{E}_f\big|\breve{f}^{(\kappa)}(y) - f(y)\big|^r \mathbb{I}_{\hat{k}\le\kappa, G_\delta^{h_{\hat{k}}}} \le \bar{C}_r D_b^r \lambda^{-r/2} S_n^r(\kappa).$$

Noting that the right hand side of the obtain inequality is independent of f and taking into account the definition of κ and h^* we obtain

$$\limsup_{n\to\infty} \sup_{f\in\mathbb{H}_d(\beta,L,A,M)} \phi_n^{-r}(\beta) R_1(f) < \infty. \qquad (5.4.17)$$

Now, let us bounded from above R_2. Applying Cauchy-Schwarz inequality we have in view of Lemma 17

$$\begin{aligned}
R_2(f) &= \sum_{k=\kappa}^{\mathbf{k}_n} \mathbb{E}_f\big|\breve{f}^{(k)}(y) - f(y)\big|^r \mathbb{I}_{G_\delta^{h_k}} \\
&\le \sum_{k>\kappa} \left(\mathbb{E}_f\big|\breve{f}^{(k)}(y) - f(y)\big|^{2r} \mathbb{I}_{G_\delta^{h_k}}\right)^{1/2} \sqrt{\mathbb{P}_f\{\hat{k} = k, G_\delta^{h_k}\}} \\
&= \sqrt{J_2} \sum_{k>\kappa} \left(\mathbb{E}_f\big|\breve{f}^{(k)}(y) - f(y)\big|^{2r} \mathbb{I}_{G_\delta^{h_k}}\right)^{1/2} 2^{-(k-1)rd}. \qquad (5.4.18)
\end{aligned}$$

We obtain from (5.4.15) and (5.4.18)

$$R_2(f) \le N_{h_{\max}}^{-r} \times D_b \frac{\bar{C}_{2r}^r\, 2^{rd}\sqrt{J_2}}{\left(\sqrt{\lambda}\right)^r} \sum_{s\ge 0} (1 + k\ln 2)^{r/2} 2^{-srd}. \qquad (5.4.19)$$

It remains to note that the right hand side of (5.4.19) is independent of f. Thus, we have

$$\limsup_{n\to\infty} \sup_{f\in\mathbb{H}_d(\beta,L,A,M)} \phi_n^{-r}(\beta) R_2(f) < \infty.$$

that yields together with (5.4.16) and (5.4.17)

$$\limsup_{n\to\infty} \sup_{f\in\mathbb{H}_d(\beta,L,A,M)} \phi_n^{-r}(\beta) \mathbb{E}_f \big|\check{f}^{(\hat{k})}(y) - f(y)\big|^r \mathbb{I}_{G_\delta^{h_{\hat{k}}}} < \infty.$$

To get the assertion of the theorem it suffices to show that

$$\limsup_{n\to\infty} \sup_{f\in\mathbb{H}_d(\beta,L,A,M)} \phi_n^{-r}(\beta) \mathbb{E}_f \big|\check{f}^{(\hat{k})}(y) - f(y)\big|^r \mathbb{I}_{\left(G_\delta^{h_{\hat{k}}}\right)^c} < \infty. \tag{5.4.20}$$

Note that $|\check{f}^{(\hat{k})}(y)| \le M$ by definition. This allows us to state that $\big|\check{f}^{(\hat{k})}(y) - f(y)\big| \le 2M$. Here we also have taken into account that $||f||_\infty \le M$.

Finally we obtain

$$R_3(f) \le 2^r \ M^r \mathbb{P}_f\big\{\big(G_\delta^{h_{\hat{k}}}\big)^c\big\}.$$

and (5.4.20) follows now from Lemma 16. ∎

5.5 Appendix

Proof of lemma 13. **1)** Remember that for any $\vec{p},\vec{q}\in\mathcal{S}_b^2$

$$\nabla_h^{\vec{p},\vec{q}}(\theta)=\frac{1}{h^d}\int_{V_h(y)}\Big(\frac{x-y}{h}\Big)^{\vec{p}+\vec{q}}\int_{-1}^{1}g_\xi\big(z\big)dz\,dx=\int_{-1}^{1}g_\xi\big(z\big)dz\int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d}x^{\vec{p}+\vec{q}}\,dx.$$

So we can write that

$$J_D(\theta)=\int_{-1}^{1}g_\xi\big(z\big)dz\int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d}K(x)\,K^\top(x)\,dx.$$

According to *Assumptions 2*, we have that $\int_{-1}^{1}g_\xi\big(z\big)dz>0$. Now we show that $\int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d}K(x)\,K^\top(x)\,dx$ is symmetric definite matrix, indeed for any $\tau\in\mathbb{R}^{D_b}\backslash 0$

$$\tau^\top\int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d}K(x)\,K^\top(x)\,dx\,\tau=\int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d}\big[\tau^\top K(x)\big]^2\,dx>0.$$

As q' is almost continuous everywhere and by definition of $J_D(.)$ in (5.4.9), it exists a neighborhood of θ noted $\mathcal{V}$ such that the matrix $J_D\big(\tilde{\theta}\big)$ is invertible for all $\tilde{\theta}\in\mathcal{V}$.

2) With *Assumptions 2*, in particular that g_ξ is even, by definition of D_h in (5.4.6) we have that $D_h(\theta)=0$. According to *1)*, the function D_h is locally invertible around θ, so $D_h(\theta)=0$ is unique solution on $\mathcal{V}$. Then, we can see that each partial derivative (5.4.8) is positive, thus θ is the unique solution of $D_h(.)=(0,...,0)$ on $\mathbb{R}^{D_b}$.

3) We can write the following equalities, for any $\tilde{\theta}\in\mathcal{V}$ then $\exists\bar{\theta}\in\mathcal{V}$ such that $D_h\big(\tilde{\theta}\big)-D_h(\theta)=J_D\big(\bar{\theta}\big)\big(\tilde{\theta}-\theta\big)$ and

$$\|D_h\big(\tilde{\theta}\big)-D_h(\theta)\|_2^2=\big(\tilde{\theta}-\theta\big)^\top J_D\big(\bar{\theta}\big)^\top J_D\big(\bar{\theta}\big)\big(\tilde{\theta}-\theta\big).$$

It is elementary to show that $J_D\big(\bar{\theta}\big)^\top J_D\big(\bar{\theta}\big)$ is definite positive matrix if $J_D\big(\bar{\theta}\big)$ is invertible. Indeed, for $\tau\in\mathbb{R}^{D_b}\backslash 0$ such that

$$\tau^\top J_D\big(\bar{\theta}\big)^\top J_D\big(\bar{\theta}\big)\tau=0\Rightarrow J_D\big(\bar{\theta}\big)\tau=0\Rightarrow\tau=0,$$

it is a contradiction. Thus with *1)*

$$\|D_h\big(\tilde{\theta}\big)-D_h(\theta)\|_2\geq\sqrt{\lambda}\,\|\tilde{\theta}-\theta\|_2,\quad\tilde{\theta}\in\mathcal{V},\tag{5.5.1}$$

where $\lambda=\inf_{\bar{\theta}\in\mathcal{V}}\lambda_{\bar{\theta}}$ and $\lambda_{\bar{\theta}}$ is the smallest eigenvalue of matrix $J_D\big(\bar{\theta}\big)^\top J_D\big(\bar{\theta}\big)$. It follows that $\lambda>0$. ∎

Proof of lemma 14. By definition of $D_h^{\vec{p}}$ (5.4.6) and $\mathcal{E}_h^{\vec{p}}$ (5.4.5), $\Delta_h^{\vec{p}}$ (5.4.7) can be written as

$$\begin{aligned}\Delta_h^{\vec{p}}(t) &= -\frac{1}{h^d}\int_{V_y(h)}\Big(\frac{x-y}{h}\Big)^{\vec{p}}\Bigg[\int_{1+f_t(x)-f(x)}^{1+f_{t-\theta}(x)} g_\xi(z)dz - \int_{-1+f_{t-\theta}(x)}^{-1+f_t(x)-f(x)} g_\xi(z)dz \\ &+ \int_{1+f_{t-\theta}(x)}^{1+f_t(x)-f(x)} \big(z-f_t(x)+f(x)\big)g_\xi(z)dz \\ &+ \int_{-1+f_t(x)-f(x)}^{-1+f_{t-\theta}(x)} \big(z-f_t(x)+f(x)\big)g_\xi(z)dz \\ &+ \int_{-1+f_{t-\theta}(x)}^{1+f_{t-\theta}(x)} \big(f(x)-f_\theta(x)\big)g_\xi(z)dz\Bigg]\,dx.\end{aligned}$$

With simple calculations, we bound the *bias term*

$$|\Delta_h^{\vec{p}}(t)| \le b_h.$$

∎

Proof of lemma 15. First, let us bound the following term for any $\vec{p} \in \mathcal{S}_b$

$$\sup_{t\in\Theta(M)}\Big|\tilde{D}_h^{\vec{p}}(t) - D_h^{\vec{p}}(t)\Big| \le \sup_{t\in\Theta(M)}\Big|\tilde{D}_h^{\vec{p}}(t) - \mathcal{E}_h^{\vec{p}}(t)\Big| + \sup_{t\in\Theta(M)}\Big|\Delta_h^{\vec{p}}(t)\Big|,$$

where $\mathcal{E}_h^{\vec{p}}$ and $\Delta_h^{\vec{p}}$ are respectively defined in (5.4.5) and (5.4.7). In view of Lemma 14, we get

$$\sup_{t\in\Theta(M)}\Big|\tilde{D}_h^{\vec{p}}(t) - D_h^{\vec{p}}(t)\Big| \le \sup_{t\in\Theta(M)}\Big|\tilde{D}_h^{\vec{p}}(t) - \mathcal{E}_h^{\vec{p}}(t)\Big| + b_h. \tag{5.5.2}$$

Note the function $L(.) = \tilde{D}_h^{\vec{p}}(.) - \mathcal{E}_h^{\vec{p}}(.)$. To show the assertion of the lemma, we use a chaining argument on $L(.)$. Remember that $\Theta(M)$ is a compact of $\mathbb{R}^{D_b}$. Let $t_0 \in \Theta(M)$ fixed and for any $l \in \mathbb{N}^*$ put Γ_l a a^{-l}-net on $\Theta(M)$ where the constant $a = \exp\Big\{\frac{4\pi^{16}}{e18^42}\Big\}$. Introduce the following notations

$$u_0(t) = t_0, \quad u_l(t) = \arg\inf_{u\in\Gamma_l}\|u-t\|_1, \quad l \in \mathbb{N}^*.$$

The chaining argument rests on the following decomposition:

$$L(t) = L(t_0) + \sum_{l=1}^{\infty} L\big(u_l(t)\big) - L\big(u_{l-1}(t)\big), \quad \forall t \in \Theta(M). \tag{5.5.3}$$

Using (5.5.2) and (5.5.3), we obtain

$$\mathbb{P}_f\left(N_h \sup_{t\in\Theta(M)} |L(t)| \geq z - b_h\, N_h\right)$$

$$\leq \mathbb{P}_f\left(|L(t_0)| + \sup_{t\in\Theta(M)} \sum_{l=1}^{\infty} \big|L\big(u_l(t)\big) - L\big(u_{l-1}(t)\big)\big| \geq \frac{z}{N_h} - b_h\right). \tag{5.5.4}$$

We can control the second term as follows.

$$\sup_{t\in\Theta(M)} \sum_{l=1}^{\infty} \big|L\big(u_l(t)\big) - L\big(u_{l-1}(t)\big)\big| \leq \sum_{l=1}^{\infty} \sup_{\substack{u,v\in\Gamma_l\times\Gamma_{l-1}\\ \|u-v\|_1\leq a^{-l}}} |L(u) - L(v)|\,,$$

where $\Gamma_0 = \{t_0\}$.Using (5.5.4) and last inequality, we get

$$\mathbb{P}_f\left(N_h \sup_{t\in\Theta(M)} |L(t)| \geq z - b_h\, N_h\right)$$

$$\leq \mathbb{P}_f\big(N_h\, |L(t_0)| \geq z/2 - b_h\, N_h/2\big)$$

$$+\mathbb{P}_f\left(N_h \sum_{l=1}^{\infty} \sup_{\substack{u,v\in\Gamma_l\times\Gamma_{l-1}\\ \|u-v\|_1\leq a^{-l}}} \big|L(u) - L(v)\big| \geq z/2 - b_h\, N_h/2\right). \tag{5.5.5}$$

We define the function $\mathcal{W}_t(x,z) = \frac{1}{N_h}\, q\big(z + f(x) - f_t(x)\big)\, \big(\frac{x-y}{h}\big)^{\vec{p}}\, \mathbb{I}_{x\in V_h(y)}$, then the process $N_h\, L(.)$ can be written as an empirical process (sum of independent zero-mean random variables)

$$N_h\, L(t) = \sum_{i=1}^{n} \mathcal{W}_t\big(X_i,\xi_i\big) - \mathbb{E}_f \mathcal{W}_t\big(X_i,\xi_i\big), \quad t\in\Theta(M) \tag{5.5.6}$$

At a fixed point t_0, we can use classical exponential inequalities for empirical process. If the design is deterministic, Hoeffding inequality is sufficient. But the random design requires the Bernstein inequality (see Boucheron, Bousquet, and Lugosi [2004], for example) which takes into account the variance of random variables. By definition of $\mathcal{W}_t(.,.)$ (in (5.5.6)), we have

$$\sum_{i=1}^{n} \mathbb{E}_f \mathcal{W}_t^2(X_i,\xi_i) \leq 1, \qquad \|\mathcal{W}_t(.,.)\|_\infty \leq 1/N_h, \tag{5.5.7}$$

where $\|.\|_\infty$ is the sup-norm.

For the first term of (5.5.5), we use Bernstein inequality. Indeed, we are in the presence of independent random variables bounded by $1/N_h$ (see definition of $\tilde{D}_h^{\vec{p}}$ in (5.4.4)) and with finite variance (see 5.5.7), then

$$\mathbb{P}_f\left(N_h\left|\tilde{D}_h^{\vec{p}}(t_0)-\mathcal{E}_h^{\vec{p}}(t_0)\right|\geq z/2-b_h\,N_h/2\right)\leq 2\exp\left\{-\frac{(z-b_h\,N_h)^2}{8+\frac{8}{3N_h}(z-b_h\,N_h)}\right\}. \tag{5.5.8}$$

Secondly, in view of (5.5.5), we write the probability with the following form,

$$\begin{aligned}&\mathbb{P}_f\left(N_h\sum_{l=1}^{\infty}\sup_{\substack{u,v\in\Gamma_l\times\Gamma_{l-1}\\ \|u-v\|_1\leq a^{-l}}}\left|L(u)-L(v)\right|\geq z/2-b_h\,N_h/2\right)\\ &\leq\sum_{l=1}^{\infty}\sum_{\substack{u,v\in\Gamma_l\times\Gamma_{l-1}\\ \|u-v\|_1\leq a^{-l}}}\mathbb{P}_f\left(N_h\,\frac{\pi^2}{6}l^2\left|L(u)-L(v)\right|\geq z/2-b_h\,N_h/2\right).\end{aligned}$$

Recall that

$$N_h\big[L(u)-L(v)\big]=\sum_{i=1}^{n}\mathcal{W}_u\big(X_i,\xi_i\big)-\mathcal{W}_v\big(X_i,\xi_i\big)-\mathbb{E}_f\left[\mathcal{W}_u\big(X_i,\xi_i\big)-\mathcal{W}_v\big(X_i,\xi_i\big)\right],$$

then we have a sum of independent centred random variables with finite variance and bounded. The main property of Huber function is that its derivative q is Lipshitz, then we have the following assertions.

$$\sum_{i=1}^{n}\mathbb{E}_f\left[\mathcal{W}_u(X_i,\xi_i)-\mathcal{W}_v(X_i,\xi_i)\right]^2\leq\|u-v\|_1^2,\quad\|\mathcal{W}_u(.,.)-\mathcal{W}_v(.,.)\|_\infty\leq\|u-v\|_1/N_h, \tag{5.5.9}$$

Using Bernstein inequality and last inequalities, we obtain

$$\begin{aligned}&\mathbb{P}_f\left(N_h\sum_{l=1}^{\infty}\sup_{\substack{u,v\in\Gamma_l\times\Gamma_{l-1}\\ \|u-v\|_1\leq a^{-l}}}\left|L(u)-L(v)\right|\geq z/2-b_h\,N_h/2\right)\\ &\leq 2\sum_{l=1}^{\infty}\sum_{\substack{u,v\in\Gamma_l\times\Gamma_{l-1}\\ \|u-v\|_1\leq a^{-l}}}\exp\left\{-\frac{36\|u-v\|_1^{-1}}{\pi^4l^4}\,\frac{(z-b_h\,N_h)^2}{8\|u-v\|_1+\frac{8}{3N_h}\,(z-b_h\,N_h)}\right\}\\ &\leq 2\sum_{l=1}^{\infty}\#(\Gamma_l)\,\#(\Gamma_{l-1})\,\exp\left\{-\frac{36\,a^l}{\pi^4l^4}\,\frac{(z-b_h\,N_h)^2}{8+\frac{8}{3N_h}\,(z-b_h\,N_h)}\right\},\end{aligned} \tag{5.5.10}$$

where $\#(\Gamma_l)$ is the cardinal of Γ_l. Recall that $\min_{l>0}\frac{a^l}{l^4}=1$, $z\geq(1\vee b_h\,N_h)$ and $\Sigma=2+2\sum_{l=1}^{\infty}\#(\Gamma_l)\,\#(\Gamma_{l-1})\,\exp\left\{-\frac{18\,a^l}{\pi^4l^4}\right\}<\infty$, using (5.5.4), (5.5.5), (5.5.8) and the last

inequality, we have

$$\mathbb{P}_f\left(N_h \sup_{t\in\Theta(M)}\left|\tilde{D}_h^{\vec{p}}(t)-\mathcal{E}_h^{\vec{p}}(t)\right|\geq z\right)\leq \Sigma\exp\left\{-\frac{(z-b_h\,N_h)^2}{8+\frac{8}{3N_h}z}\right\}.$$

The lemma is proved. ∎

Proof of Lemma 16. Remember that the event $\left(G_\delta^h\right)^c$ can be written as $\left(G_\delta^h\right)^c=\left\{\breve{\theta}(h)\notin\mathcal{B}(\theta,\delta)\right\}$ and $\breve{\theta}(h)$ and θ are respectively the unique solutions of equations $\tilde{D}_h\big(\breve{\theta}(h)\big)=0$ and $D_h(\theta)=0$. We can remark the following inclusion

$$\left\{\breve{\theta}(h)\notin\mathcal{B}(\theta,\delta)\right\}\subseteq\left\{\sup_{t\in\Theta(M)\setminus\mathcal{B}(\theta,\delta)}\|\tilde{D}_h(t)-D_h(t)\|_2\geq\varkappa_\delta\right\}.$$

where $\varkappa_\delta=\inf_{t\in\Theta(M)\setminus\mathcal{B}(\theta,\delta)}\frac{\|D_h(t)\|_2}{2}$. In view of Lemma 13, $\varkappa_\delta$ is strictly positive and does not depend on n. Applying the last inclusion, we obtain

$$\mathbb{P}_f\left(\left(G_\delta^h\right)^c\right)\leq\sum_{\vec{p}\in\mathcal{S}_b}\mathbb{P}_f\left(N_h\sup_{t\in\Theta(M)\setminus\mathcal{B}(\theta,\delta)}\left|\tilde{D}_h^{\vec{p}}(t)-D_h^{\vec{p}}(t)\right|>\frac{N_h\varkappa_\delta}{\sqrt{D_b}}\right)$$

The assumptions on n,h in Lemma 16 allow to show that

$$\frac{N_h\varkappa_\delta}{2\sqrt{D_b}}\geq\big(1\vee b_h\,N_h\big).$$

Using Lemma 15, we have

$$\mathbb{P}_f\left(\left(G_\delta^h\right)^c\right)\leq D_b\Sigma\,\exp\left\{-\frac{nh^d\varkappa_\delta^2}{4D_b\left(8+4\varkappa_\delta/(3\sqrt{D_b})\right)}\right\}.$$

The lemma is proved. ∎

Proof of Lemma 17. Note that for any $k\geq\kappa+1$ and by definition of $\hat{k}$ (5.3.3)

$$\left\{\hat{k}=k\right\}=\cup_{l\geq k}\left\{\left|\breve{f}^{(k-1)}(y)-\breve{f}^{(l)}(y)\right|>C\,S_n(l)\right\}.$$

Note that $S_n(l)$ is monotonically increasing in l and, therefore,

$$\begin{aligned}\left\{\hat{k}=k\right\}\;\subseteq\;&\left\{\left|\breve{f}^{(k-1)}(y)-f(y)\right|>2^{-1}C\,S_n(k-1)\right\}\\&\cup\left[\cup_{l\geq k}\left\{\left|\breve{f}^{(l)}(y)-f(y)\right|>2^{-1}C\,S_n(l)\right\}\right].\end{aligned}$$

We come to the following inequality: for any $k \geq \kappa + 1$

$$\begin{aligned} \mathbb{P}\left(\hat{k} = k, G_\delta^{h_k}\right) &\leq \mathbb{P}\left\{\left|\breve{f}^{(k-1)}(y) - \breve{f}(y)\right| > 2^{-1} C\, S_n(k-1), G_\delta^{h_k}\right\} \\ &\quad + \sum_{l \geq k} \mathbb{P}\left\{\left|\breve{f}^{(l)}(y) - f(y)\right| > 2^{-1} C\, S_n(l), G_\delta^{h_k}\right\}. \end{aligned} \tag{5.5.11}$$

Note that the definition of $S_n(l)$ yields

$$N_{h_l}\, S_n(l) \geq \left[1 + \ln\left(h_{\max}/h_l\right)\right]^{1/2}.$$

Thus, applying Proposition 7 with $\varepsilon = C\left[1 + \ln\left(h_{\max}/h_l\right)\right]^{1/2}$, we obtain $\forall l \geq k-1$

$$\begin{aligned} \mathbb{P}\left\{\left|\breve{f}^{(l)}(y) - f(y)\right| > 2^{-1} C\, S_n(l), G_\delta^{h_k}\right\} &\leq D_b \Sigma \left[h_{\max}/h_l\right]^{-2rd} \\ &= D_b \Sigma\, 2^{-2rd\, l}. \end{aligned} \tag{5.5.12}$$

Here, we have also used that $k \geq \kappa + 1$. We obtain from (5.5.11) and (5.5.12) that $k \geq \kappa + 1$

$$\mathbb{P}\left(\hat{k} = k, G_\delta^{h_k}\right) \leq J_2 2^{-2(k-1)rd},$$

where $J_2 = D_b \Sigma\left(1 + (1 - 2^{-2rd})^{-1}\right)$. ∎

Bibliography

Akaike H. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973. cf. page(s): 37

Anscombe F.J. The transformation of poisson, binomial and negative-binomial data. *Biometrika*, 35:246–254, 1948. cf. page(s): 21, 59, 74

Arias-Castro E. et Donoho D. L. Does median filtering truly preserve edges better than linear filtering? *Ann. Statist.*, 37(3):1172–1206, 2009. cf. page(s): 32, 82, 154

Astola J., Egiazarian K., et Katkovnik V. Adaptive window size image de-noising based on intersection of confidence intervals (ici) rule. *J. Math. Imaging Vis.*, 16(3):223–235, 2002. ISSN 0924-9907. cf. page(s): 40

Astola J., Egiazarian K., Foi A., et Katkovnik V. From local kernel to nonlocal multiple-model image denoising. *Int. J. Comput. Vision*, 86(1):1–32, 2010. cf. page(s): 40, 82, 154

Aubert G. et Aujol J. A variational approach to removing multiplicative noise. *SIAM J. Appl. Math.*, 68(4):925–946, 2008. cf. page(s): 114

Autin F. *Point de vue Maxiset en estimation non paramétrique*. PhD thesis, Université Paris 7, 2004. cf. page(s): 17, 37, 39

Autin F. Maxiset for density estimation on $\mathbb{R}$. *Math. Methods Statist.*, 15(2):123–145, 2006. cf. page(s): 34

Autin F., Picard D., et Rivoirard V. Large variance Gaussian priors in Bayesian nonparametric estimation: a maxiset approach. *Math. Methods Statist.*, 15(4):349–373, 2006. cf. page(s): 34

Autin Florent. Maxisets for μ-thresholding rules. *TEST*, 17(2):332–349, 2008. cf. page(s): 34

Barron A., Birgé L., et Massart P. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999. cf. page(s): 34

Bergh J. et Löfström J. *Interpolation spaces. An introduction*. Springer-Verlag, Berlin, 1976. Grundlehren der Mathematischen Wissenschaften, No. 223. cf. page(s): 17

Bhattachar S. et Sundareshan M.K. Super-resolution of sar images using bayesian and convex set-theoretic approaches. *SPIE proceding series*, 4053:189–198, 2000. cf. page(s): 119

Birgé L. et Massart P. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001. cf. page(s): 37

Birgé L. et Massart P. Minimum contrast estimators on sieves : exponential bounds and rate of convergences. *Bernoulli*, 4(3):329–375, 1998. cf. page(s): 15

Borovkov A. Statistique mathématique. *Mir, Moscou*, 1987. cf. page(s): 14

Boucheron S., Bousquet O., et Lugosi G. Concentration inequalities. *Springer*, 2004. cf. page(s): 15, 50, 51, 63, 151, 171

Bousquet O. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002. cf. page(s): 15, 66

Brown L. et Low M. Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24(6):2384–2398, 1996. cf. page(s): 25, 35, 42

Brown L.D., Cai T. Tony, et Zhou H.H. Robust nonparametric estimation via wavelet median regression. *Ann. Statist.*, 36(5):2055–2084, 2008. cf. page(s): 15, 19, 61, 154

Buades A., Coll B., et Morel J-M. A review of image denoising algorithms, with a new one. *Multiscale Model. Simul.*, 4(2):490–530, 2005. cf. page(s): 40

Bunea F., Tsybakov A., et Wegkamp M. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194 (electronic), 2007. cf. page(s): 36

Cai T. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.*, 27(3):898–924, 1999. cf. page(s): 19, 61, 154

Cavalier L. Nonparametric statistical inverse problems. *Inverse Problems*, 24(3):034004, 19, 2008. cf. page(s): 25

Cavalier L. et Golubev Y. Risk hull method and regularization by projections of ill-posed inverse problems. *Ann. Statist.*, 34(4):1653–1677, 2006. cf. page(s): 25, 38

Cavalier L. et Tsybakov A. B. Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields*, 123(3):323–354, 2002. cf. page(s): 25

Cavalier L., Golubev G. K., Picard D., et Tsybakov A. B. Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3):843–874, 2002. Dedicated to the memory of Lucien Le Cam. cf. page(s): 25

Chang X.W. et Guo Y. Huber's m-estimation in relative gps positioning: computational aspects. *Journal of Geodesy.*, 2005. cf. page(s): 32, 154

Chichignoud M. Minimax and minimax adaptive estimation in multiplicative regression : locally bayesian approach. *Revision*, 2010a. cf. page(s): 34, 35, 42, 55, 56, 85, 113, 156

Chichignoud M. Minimax adaptive estimation via locally bayesian approach. *Preprint*, 2010b. cf. page(s): 35, 42, 55, 85

Chichignoud M. Adaptive huber m-estimation in nonparametric regression. *Preprint*, 2010c. cf. page(s): 35, 42, 153

DeVore R. et Lorentz G. *Constructive approximation*, volume 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993. cf. page(s): 17

Donoho D., Johnstone I., Kerkyacharian G., et Picard D. Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, 57(2):301–369, 1995. With discussion and a reply by the authors. cf. page(s): 14, 35, 37, 39, 42

Donoho D. L. et Johnstone I. M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. cf. page(s): 34, 37

Efromovich S. Yu. et Pinsker M. S. A self-training algorithm for nonparametric filtering. *Avtomat. i Telemekh.*, (11):58–65, 1984. cf. page(s): 15, 33, 34, 35, 38, 42

Ghosal Subhashis, Lember Jüri, et Van der Vaart Aad. Nonparametric Bayesian model selection and averaging. *Electron. J. Stat.*, 2:63–89, 2008. cf. page(s): 54, 81

Goldenshluger A. et Lepski O.V. Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(3):1150–1190, 2008. cf. page(s): 25, 35, 39, 42, 51, 81, 115, 155, 158

Goldenshluger A. et Lepski O.V. Structural adaptation via lp-norm oracle inequalities. *Probab. Theory and Related Fields*, 143:41–71, 2009a. cf. page(s): 25, 39, 41, 51, 80, 81, 115, 155

Goldenshluger A. et Lepski O.V. Uniform bounds for norms of sums of independent random functions. 2009b. cf. page(s): 15, 51, 155

Goldenshluger A. et Nemirovski A. On spatially adaptive estimation of nonparametric regression. *Math. Methods Statist.*, 6(2):135–170, 1997. cf. page(s): 35, 39, 42, 47

Golubev Y. et Spokoiny V. Exponential bounds for minimum contrast estimators. *Electron. J. Stat.*, 3:712–746, 2009. cf. page(s): 15

Hall Peter et Jones M. C. Adaptive M-estimation in nonparametric regression. *Ann. Statist.*, 18(4):1712–1728, 1990. cf. page(s): 19, 30, 61, 154

Härdle W. et Tsybakov A.B. Robust nonparametric regression with simultaneous scale curve estimation. *Ann. Statist.*, 16(1):120–135, 1988. cf. page(s): 30, 154

Härdle W. et Tsybakov A.B. Robust locally adaptive nonparametric regression. In *Data analysis and statistical inference*, pages 127–144. Eul, Bergisch Gladbach, 1992. cf. page(s): 19, 30, 61, 154

Härdle W., Kerkyacharian G., Picard D., et Tsybakov A. B. *Wavelets, approximation, and statistical applications*, volume 129 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1998. ISBN 0-387-98453-4. cf. page(s): 14, 17

Härlde W. et Tsybakov A.B. Local polynomial of the volatility function in nonparametric autoregression. *J. Econometrics*, 1:223–242, 1997. cf. page(s): 114

Has'minskii R.Z. et Ibragimov I.A. *Statistical Estimation, Asymptotic Theory*. Springer-Verlag, Applications of Mathematics, 1981. cf. page(s): 18, 25, 28, 29, 30, 33, 55, 57, 60, 79, 89, 95, 109, 115, 119, 127, 131

Huber P. et Ronchetti E. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., Hoboken, NJ, second edition, 2009. cf. page(s): 31, 154

Huber P.J. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101, 1964. cf. page(s): 32, 154

Huber P.J. *Robust statistics*. Wiley, New York., 1981. cf. page(s): 32, 61, 154

Härlde W., Hart J., Marron J.S., et Tsybakov A.B. Bandwidth choice for average derivative estimation. *journal of the American Statistical Association*, 87:417, 1992. cf. page(s): 151

Juditsky A. Wavelet estimators: adapting to unknown smoothness. *Math. Methods Statist.*, 6(1):1–25, 1997. cf. page(s): 35, 42

Juditsky A.B., Lepski O.V., et Tsybakov A.B. Nonparametric estimation of composite functions. *Ann. Statist.*, 37(3):1360–1404, 2009. cf. page(s): 115, 155

Katkovnik V. *Nonparametric identification and data smoothing*. "Nauka", Moscow (in Russian), 1985. The method of local approximation. cf. page(s): 26

Katkovnik V. A new method for varying adaptive bandwidth selection. *IEEE Trans. Image Process.*, 47(9):2567–2571, 1999. cf. page(s): 39, 40, 47, 48

Katkovnik V. et Spokoiny V. Spatially adaptive estimation via fitted local likelihood techniques. *IEEE Trans. Image Process.*, 56(3):873 –886, 2008. cf. page(s): 19, 29, 39, 87, 115, 117, 119, 156

Kerkyacharian G. et Picard D. Minimax or maxisets? *Bernoulli*, 8(2):219–253, 2002. cf. page(s): 34

Kerkyacharian G., Lepski O.V., et Picard D. Non linear estimation in anisotropic multiindex denoising. *Probab. Theory and Related Fields*, 121:137–170, 2001. cf. page(s): 25, 39, 51, 80, 81, 155, 158

Kervrann Ch. et Boulanger J. Optimal spatial adaptation for patch-based image denoising. *IEEE Trans. Image Process.*, 15(10):2866–2878, 2006. cf. page(s): 40

Klutchnikoff N. *On the adaptive estimation of anisotropic functions*. PhD thesis, Aix-Masrseille 1, 2005. cf. page(s): 25, 34, 35, 39, 42, 51, 56, 57, 60, 62, 81, 82, 88, 94, 95, 118, 123, 125, 155, 158, 160

Korostelëv A. P. et Tsybakov A. B. *Minimax theory of image reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1993. ISBN 0-387-94028-6. cf. page(s): 114

Lecué G. et Mendelson S. Aggregation via empirical risk minimization. *Probab. Theory Related Fields*, 145(3-4):591–613, 2009. cf. page(s): 37

Ledoux M. On talagrand's deviation inequalities product measures. *ESAIM: Probability and Statistic*, 1:63–87, 1997. cf. page(s): 15, 51, 66

Lepski O. V., Mammen E., et Spokoiny V. G. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.*, 25(3):929–947, 1997. ISSN 0090-5364. cf. page(s): 35, 39, 41, 42, 51, 88, 117, 123, 137, 157, 158, 166

Lepski O.V. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability and its Applications*, 35(3):454–466, 1990. cf. page(s): 33, 34, 35, 37, 38, 41, 42, 56, 88, 89, 117, 118, 119, 157, 158, 160

Lepski O.V. Asymptotically minimax adaptive estimation i. upper bounds. optimally adaptive estimates. *Theory Probab. Appl.*, 36:682–697, 1991. cf. page(s): 35, 38, 41, 42, 88, 117, 121, 123, 137, 157

Lepski O.V. Asymptotically minimax adaptive estimation ii. statistical models without optimal adaptation. adaptive estimators. *Theory of Probability and its Applications*, 37: 433–468, 1992a. cf. page(s): 35, 39, 42, 89, 119, 121, 160

Lepski O.V. On problems of adaptive estimation in white gaussian noise. *In Topic in Nonparametric Estimation*, 12:87–106, 1992b. cf. page(s): 35, 39, 42, 160

Lepski O.V. et Levit B.Y. Adaptive nonparametric estimation of smooth multivariate functions. *Mathematicals methods of statistics*, 1999. cf. page(s): 39

Lepski O.V. et Spokoiny V.G. Optimal pointwise adaptive methods in nonparametric estimation. *Annals of statistics*, 25(6):2512–2546, 1997. cf. page(s): 35, 42, 88, 118, 157, 160

Mallows C. Some comments on c_p. *Technometrics.*, 15:661–675, 1973. cf. page(s): 37

Marteau C. On the stability of the risk hull method for projection estimators. *J. Statist. Plann. Inference*, 139(6):1821–1835, 2009. cf. page(s): 25, 38

Massart P. Some applications of concentration inequalities to statistics. *Probability Theory*, volume spécial dédié à Michel Talagrand(2):245–303, 2000. cf. page(s): 15, 51, 66

Massart P. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. cf. page(s): 15, 35, 51, 66, 67

Mathé P. The Lepskiĭ principle revisited. *Inverse Problems*, 22(3):L11–L15, 2006. cf. page(s): 39

Meyer Y. *Wavelets and operators*, volume 37 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1992. Translated from the 1990 French original by D. H. Salinger. cf. page(s): 17, 37

Nadaraya E.A. On estimating regression. *Theory of Probability and its Applications*, 9(1): 141–142, 1964. cf. page(s): 14

Nemirovski A. *Topics in non-parametric statistics*, volume 1738 of *Lecture Notes in Math.* Springer, Berlin, 2000. cf. page(s): 20, 27, 37, 147

Nussbaum M. Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.*, 24(6):2399–2430, 1996. cf. page(s): 25

Parzen E. On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962. cf. page(s): 14

Peetre J. *New thoughts on Besov spaces*. Mathematics Department, Duke University, Durham, N.C., 1976. Duke University Mathematics Series, No. 1. cf. page(s): 17

Petrus P. Robust huber adaptive filter. *IEEE Transactions on Signal Processing.*, 47:1129–1133, 1999. cf. page(s): 32, 154

Plancade Sandra. Nonparametric estimation of the density of the regression noise. *C. R. Math. Acad. Sci. Paris*, 346(7-8):461–466, 2008. cf. page(s): 36

Polzehl J. et Spokoiny V. Adaptive weights smoothing with applications to image restoration. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 62(2):335–354, 2000. ISSN 1369-7412. cf. page(s): 39

Polzehl J. et Spokoiny V. Image denoising: pointwise adaptive approach. *Ann. Statist.*, 31 (1):30–57, 2003. cf. page(s): 39

Polzehl J. et Spokoiny V. Propagation-separation approach for local likelihood estimation. *Probab. Theory Related Fields*, 135(3):335–362, 2006. cf. page(s): 19, 29, 30, 39, 87, 115, 117, 119

Reiss M., Rozenholc Y., et Cuenod C. Pointwise adaptive estimation for robust and quantile regression. 2009. Source: Arxiv. cf. page(s): 19, 30, 61, 81, 155

Rigollet Ph. et Tsybakov A. B. Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280, 2007. cf. page(s): 37

Rosenblatt M. Remarks on som nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 23:832–837, 1956. cf. page(s): 14

Rousseeuw P. et Leroy A. *Robust regression and outlier detection*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1987. cf. page(s): 31, 154

Rousseeuw P.J. Least median of squares regression. *J. Amer. Statist. Assoc.*, 79(388): 871–880, 1984. cf. page(s): 154

Réfrégier P. *Théorie du bruit et applications en physique*. Hermès Science Publications, 2002. cf. page(s): 119

Simar L. et Wilson P. Statistical inference in nonparametric frontier models: The state of the art. *Journal of Productivity Analysis*, 13:49–78, 2000. cf. page(s): 24, 114

Spokoiny V. Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *Ann. Statist.*, 26(4):1356–1378, 1998. cf. page(s): 39

Spokoiny V. et Vial C. Parameter tuning in pointwise adaptation using a propagation approach. *Ann. Statist.*, 37(5B):2783–2807, 2009. cf. page(s): 39, 41

Spokoiny V. G. Adaptive hypothesis testing using wavelets. *Ann. Statist.*, 24(6):2477–2498, 1996. cf. page(s): 35, 42

Stein C. M. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9 (6):1135–1151, 1981. cf. page(s): 14, 61, 154

Stone C. Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8(6): 1348–1360, 1980. cf. page(s): 33

Stone C. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 1982. cf. page(s): 15, 33, 38

Talagrand M. Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, 22(1): 28–76, 1994. cf. page(s): 51, 66

Talagrand M. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'I.H.E.S*, 81:73–205, 1995. cf. page(s): 15, 66

Talagrand M. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996a. cf. page(s): 15, 66

Talagrand M. A new look at independence. *Annals of Probability*, 24:1–34, 1996b. cf. page(s): 15, 66

Tsybakov A. B. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008. cf. page(s): 19, 20, 26, 31, 32, 34, 82, 93, 109, 151, 159

Tsybakov A. B. Robust estimates of a function. *Problems of Information Transmission*, 18 (3):39–52, 1982b. cf. page(s): 30, 61, 154

Tsybakov A. B. Convergence of nonparametric robust algorithms of reconstruction of functions. *Automation and Remote Control*, (12):66–76, 1983. cf. page(s): 30, 61, 154

Tsybakov A. B. Robust reconstruction of functions by a local approximation method. *Problems of Information Transmission*, 22(2):69–84, 1986. cf. page(s): 30, 61, 154

Tsybakov A.B. Nonparametric signal estimation when there is incomplete information on the noise distribution. *Problems of Information Transmission*, 18(2):116–130, 1982a. cf. page(s): 30, 32, 61, 154, 159

Tsybakov A.B. Pointwise and sup-norm sharp adaptive estimation of function on the sobolev classes. *Annals of statistics*, 26(6):2420–2469, 1998. cf. page(s): 35, 42, 56

Van de Geer S. *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000. cf. page(s): 64, 154, 165

Van der Vaart A. W. et Van Zanten J. H. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675, 2009. cf. page(s): 54, 81

Van der Vaart Aad W. et Wellner Jon A. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics. cf. page(s): 67

Watson G. Smooth regression analysis. *Sankhya, Ser.A*, 26:359–372, 1964. cf. page(s): 14

Woodroofe M. On choosing a delta-sequence. *Ann. Math. Statist.*, 41:1665–1671, 1970. cf. page(s): 19

Résumé

On se place dans le cadre de l'estimation non paramétrique dans le modèle de régression. Dans un premier temps, on dispose des observations Y dont la densité g est connue et dépend d'une fonction de régression $f(X)$ inconnue. Dans cette thèse, cette fonction est supposée régulière, i.e. appartenant à une boule de Hölder. Le but est d'estimer la fonction f à un point y (estimation ponctuelle). Pour cela, nous développons un estimateur local de type *bayésien*, construit à partir de la densité g des observations. Nous proposons une procédure adaptative s'appuyant sur la méthode de Lepski, qui permet de construire un estimateur adaptatif choisi dans la famille des estimateurs bayésiens locales indexés par la fenêtre. Sous certaines hypothèses suffisantes sur la densité g, notre estimateur atteint la vitesse adaptative optimale (en un certain sens). En outre, nous constatons que dans certains modèles, l'estimateur bayésien est plus performant que les estimateurs linéaires.

Ensuite, une autre approche est considérée. Nous nous plaçons dans le modèle de régression additive, où la densité du bruit est inconnue, mais supposée symétrique. Dans ce cadre, nous développons un estimateur dit de *Huber* reposant sur l'idée de la médiane. Cet estimateur permet d'estimer la fonction de régression, quelque soit la densité du bruit additif (par exemple, densité gaussienne ou densité de Cauchy). Avec la méthode de Lepski, nous sélectionnons un estimateur qui atteint la vitesse adaptative classique des estimateurs linéaires sur les espaces de Hölder.

Mots Clés : Approche Bayésienne Locale, Critère de Huber, Estimation Robuste, Méthode de Lepski, Adaptation Minimax, Régression Multiplicative, Sélection de la Fenêtre, Vitesses Adaptatives Optimales.

Abstract

We work in the context of nonparametric estimation in the regression model. Firstly, we consider observations Y where the density g is known and depends on a regression function $f(X)$ unknown. In this thesis, this function is assumed regular, i.e. belonging to a Hölder ball. The goal is to estimate the function f to a point y (pointwise estimation). For it, we develop a *local bayesian estimator*, constructed from the density g of the observations. We propose an adaptive procedure based on the Lepski's method, which allows to construct an adaptive estimator chosen from the family of *local bayesian estimators* indexed by the bandwidth. Under some sufficient assumptions on the density g, our estimator achieves the adaptive optimal rate (in a particular sense). In addition, we remark that, in some models, the bayesian estimator is more efficient than linear estimators.

Secondly, another approach is considered. We consider the additive regression model, where the density of the noise is unknown and assumed to be symmetric. In this framework, we develop the so-called *Huber estimator* based on the idea of the median (ℓ_1 criterion). This estimator allows to estimate the regression function, for any density of the additive noise (for example : Gaussian density or Cauchy density). With the Lepski's method, we select an estimator that achieves the rate of conventional adaptive linear estimators on Hölder spaces without depends on the symmetric density of the noise.

Keywords : Bandwidth Selector, Adaptive Optimal Rate, Huber Criterion, Lepski's Method, Local Bayesian Fitting, Minimax Adaptation, Multiplicative Regression, Robust Estimation.

AMS 2000 subject classification: 62G08, 62G20.